

University of Oklahoma
Graduate College

CELL-CELL INTERACTIONS IN *Myxococcus xanthus*

A Dissertation
SUBMITTED TO THE GRADUATE FACULTY
in partial fulfillment of requirements for the
degree of
Doctor of Philosophy

By

Yanglong Zhu
Norman, Oklahoma
2004

UMI Number: 3143548



UMI Microform 3143548


Copyright 2005 by ProQuest Information and Learning Company.
All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

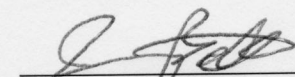
ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

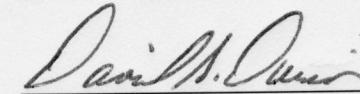
CELL-CELL INTERACTIONS IN *Myxococcus xanthus*

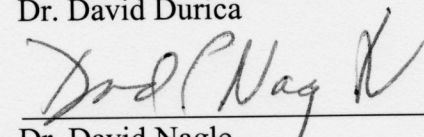
A Dissertation Approved for the
DEPARTMENT OF BOTANY AND MICROBIOLOGY

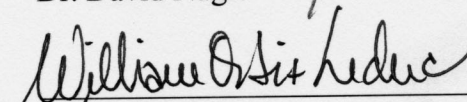
BY


Dr. John Downard


Dr. Jim Ballard


Dr. David Durica


Dr. David Nagle


Dr. William Ortiz

DEDICATION

To my wife and daughter for their unwavering support, and
To my parents for their faith in educating their children whatever it takes.

ACKNOWLEDGEMENT

I gladly acknowledge the effort made by Dr. John Downard in my graduate education at the University of Oklahoma. He provided a great latitude of freedom in the course of my research and offered unfailing help whenever I needed. His approach of graduate education has instilled in me independent thinking and creativity. I also acknowledge Dr. Downard's kindness and generosity towards my personal life.

My sincere thanks are also due to my graduate committee members: Dr. Jim Ballard, Dr. David Durica, Dr. David Nagle, and Dr. William Ortiz-Leduc for their patience with my research and generous encouragement. I also acknowledge the contribution to my graduate education made by previous committee members who have left the University of Oklahoma: Dr. Alice DeMerris and Dr. David McCarthy.

I am grateful to the faculty in the Botany and Microbiology Department for their kindness and support. In particular, I have had fun to work as a teaching assistant with many of the faculty members: Dr. Jim Ballard, Dr. Tyrrell Conway, Dr. John Downard, Dr. Lee Krumholz, Dr. John Lancaster, Dr. David Nagle. I have learned a lot from them in the role of a teaching assistant. I also thank Dr. William Ortiz-Leduc for providing me an open access to his spectrophotometer. I will never forget all those helps provided by the staff member in the department.

A very special thanks goes to my wife, Lihua Gu, for her support and concerns all these years. Finally, I thank my parents for supporting my education through out my life.

TABLE OF CONTENTS

Chapter 1: A genome-wide survey of genetic loci involved in exopolysaccharide biosynthesis in <i>Myxococcus xanthus</i> .	1
Introduction	1
Materials and methods	8
Results and discussion	19
Conclusion	96
Chapter 2: A computer program for automated insertion point mapping	105
Introduction	105
Statement of problem	119
Results	124
Discussion	136
Appendix	142
Chapter 3: A-signal processing requires the catabolic pathways of A-factors	143
Introduction	143
Materials and methods	149
Results	152
Discussion	167
References	185

LIST OF TABLES

Table	Page
1.1 Primers used for internal fragment replacement mutagenesis	12
1.2 Relative social motility	32
1.3 Fluorescence patterns in the test spot on 0.3% agar containing CYE	44
1.4 Single pass sequencing error rates (a sample).	46
1.5 The relationship between the insertions, the Rcontigs and the NCBI database contigs	52
3.1 Primers used for internal fragment replacement PCR mutagenesis	136
3.2 Relative activity of the <i>dcm-1</i> strain compared to the wildtype DK1622	150

LIST OF FIGURES

Figure	Page
1.1 The transposon magellan-4 structure as it is inserted in the genome	9
1.2 Schematic drawing to show how the internal fragment replacement mutagenesis works	11
1.3 Growth comparisons	20-23
1.4 Relative cohesion assay	26-28
1.5 Social motility assay	29-30
1.6 Development defects of the mutants	33-38
1.7 Calcofluor White binding assay	39-40
1.8 Enhanced Calcofluor white binding	41-42
1.9 Insertion cds19 maps to the Rcontig8	47
1.10 Genomic sequence map	49
1.11 Insertion map of Cluster 1	61
1.12 Alignment between the orfL1-4-29 and the transcriptional regulator domain COG5340	64
1.13 Map of insertion cds4	69
1.14 Insertion map of cds1 at Cluster 3	70
1.15 Map of insertion cds13 at Cluster 4	72
1.16 Transposon insertion sites in the Cluster 5	68
1.17 A model type IV pili system	70
1.18 Transposon insertion sites in the Cluster 6	77
1.19 Internal fragment replacement mutagenesis sites near SR53	87

1.20	Sites of internal fragment replacement mutagenesis near <i>cds29</i>	89
1.21	Cohesion comparison	91
1.22	Development assay comparing strains carrying mutation in the <i>cds29</i> operon	92
1.23	1.23 Development assay comparing strains carrying mutation near the <i>cds53</i> insertion	93
1.24	<i>M. xanthus</i> cells on Calcofluor white containing plates	94
2.1	Alignment of two DNA sequences	106
2.2	BLAST matches and alignments	108-111
2.3	Showing the occasional misalignment in BLAST output	112-113
2.4	Show the confusing part of a BLAST match	114
2.5	An example output from the SHAPE program	114
2.6	CLUSTAL W output without complementation	115-116
2.7	CLUSTAL W output with proper complementation	117-118
2.8	The structure of the transposon <i>magellan-4</i>	120
2.9	Circularized <i>magellan-4</i> plasmid	121
2.10	An example of 2 dimensional Levenshtein distance matrix	126
2.11	Insertion <i>cds19</i> maps to the Rcontig8	128
2.12	The insertion in the genome	129
2.13	The insertion is cloned in a self-cloning plasmid	129
2.14	Each flanking sequence forms two stretches of homology with the Rcontig1.	130
2.15	Layered presentation of the multiple query sequences	132

2.16	Only one side of the insertion is known	133
2.17	Duplicated sequences are clearly shown	133
2.18	When Rcontig is limited on the 5'-end	134
2.19	When the Rcontig is limited on the 3'-end	134
2.20	The reliability of mapping results is clearly visible	134
2.21	The flanking sequence closes the gap between two Rcontigs	135
3.1	Comparison of A-factor activities between the wildtype and the <i>esg</i> mutant in response to branched chain amino acids	153
3.2	Degradation pathways for the branched amino acids	154
3.3	Comparison of A-factor activities of selected amino acids between the wildtype and the <i>esg</i>	155
3.4	<i>esg</i> carries defects in responding to wildtype-conditioned medium	156
3.5	Compare A-factor activities from <i>pah</i> mutant with the wildtype in response to leucine and phenylalanine	160
3.6	Compare A-factor activities from the <i>aldA</i> strain with the wildtype in response to alanine and pyruvate	162
3.7	A survey of A-factor activity of arginine and its urea cycle degradation intermediates	163
3.8	The urea cycle and arginine degradation pathway in <i>Myxococcus xanthus</i>	164
3.9	Compare A-factor activities from the <i>dcm-1</i> strain with the wildtype in response to short chain fatty acids	166
3.10	The glyoxylate cycle	172

3.11	The lone <i>argE</i> locus at 1.55Mbp on the physical map	174
3.12	The A-signaling model	184

ABSTRACT

Myxococcus xanthus, a Gram-negative, rod-shaped, nonpathogenic soil bacterium has been used as a model organism for many years for research on developmental behavior such as multicellular morphogenesis, signal transduction. This dissertation is focused on two aspects of the development. First, since *M. xanthus* development requires polysaccharide structures such as fibrils, a genome-wide survey of polysaccharide production genes were carried out with two methods: 1) using transposon mutagenesis to produce exopolysaccharide production-deficient mutants, then clone the genomic sequences flanking the transposon insertions for sequencing; 2) searching the *M. xanthus* genomic sequence database for homologs to known polysaccharide production-related genes in other organisms. Sequence analysis indicated that the transposon mutants appear to include at least two polysaccharide export systems required for exopolysaccharide production. The evidence is: 1) that many genes at separate loci, related to the type IV pilus biogenesis system such as *pilQ* are required for exopolysaccharide production. This is consistent with the hypothesis of pil proteins forming a general export system for polysaccharides involved in development; and 2) that some open reading frames (ORFs) around the insertion *cds29* are required for exopolysaccharide production and homologous to genes such as *gumB*, *gumC* (outer membrane proteins) and *rfbX* (cytoplasmic membrane proteins). Considering an earlier report on the requirement of *rfbABC* for *M. xanthus* development, this seems consistent with the hypothesis of the gene products forming a polysaccharide-specific export system. This is the first indication that a

polysaccharide-specific export system exists in *M. xanthus*. In addition to the apparent export systems, six glycosyltransferases, at least one chemotaxis gene, one chaperone gene, and a few other ORFs were identified as being required for exopolysaccharide production. Many chemotaxis genes were found to be interspersed in the putative operons of polysaccharide production-related genes. This arrangement may be a way for *M. xanthus* to coordinate the sensory system with the polysaccharide production system involved in development. Sequence searches in the *M. xanthus* genome database found many more glycosyltransferase homologs in addition to the ones identified by the transposon insertions.

A computer program, called “SHAPE”, was developed to automate the mapping of transposon insertions to database sequences (especially when the database sequences are not available in a public database), and produce a graphical representation of the results. A second computer program was developed to generate a genomic sequence map before the whole genome is completely sequenced. This program is novel, and useful for genomic sequencing and analysis, especially for closing final gaps between the contigs.

In a third portion of this work, a study was done on one of the five putative signals (identified previously), which are required for fruiting body formation during *M. xanthus* starvation-induced development. A-signal, a set of amino acids such as phenylalanine, tyrosine, proline, tryptophan, leucine and isoleucine, is the

first signal in development. A-signal response was found to be novel in that processing of A-signal amino acids via degradation is required for producing A-signal response. It seems that *M. xanthus* degrades the A-signal amino acids both to derive carbon and energy, and to produce A-signal response. This would be the first example for a signal response to be based on the availability of carbon and energy from the signal molecules.

CHAPTER 1

A GENOME-WIDE SURVEY OF GENETIC LOCI INVOLVED IN EXOPOLYSACCHARIDE BIOSYNTHESIS IN *Myxococcus xanthus*

INTRODUCTION

Myxococcus xanthus is a Gram negative, long-rod shaped, aerobic bacterium found in soil around the world. *M. xanthus* is motile with gliding motility. It has two life cycles: vegetative cycle and developmental cycle. When nutrients are sufficient in its environment, *M. xanthus* grows vegetatively, multiplying exponentially every 4 to 5 hours. Under starvation conditions, *M. xanthus* will go through a developmental cycle, in which tens of thousands of cells move into a well-distributed, raised structure called a fruiting body, and form spores there. This chapter is a survey of the genes involved in exopolysaccharide biosynthesis, and development using transposon insertion mutagenesis and genomic sequence analysis.

Gliding motility refers to a slow surface-associated translocation in the direction of the cell's long axis. It occurs in many microorganisms, including Myxobacteria, Cyanobacteria, and Flexibacteria. *M. xanthus* has two genetically distinct gliding motility systems: social (S) motility with which cells move in groups, and adventurous (A) motility with which cells move individually, adventuring away from other cells. The force for S-motility is generated by retraction of type IV pilus (Wu *et al.*, 1997; Sun and Zusman, 2000, Sherker and Berg, 2001), which requires the chitin-like component from the extracellular polysaccharides (Li

et al., 2003). The force for A-motility is believed to be associated with the extrusion of polysaccharide from tiny polar nozzles (Wolgemuth *et al.*, 2002). Wolgemuth and colleagues believe that hydration of highly concentrated polysaccharide spouted out from the nozzles generates the motility force.

M. xanthus has structures on the cell surface called fibrils (Behmlander and Dworkin, 1991; Ramaswamy and Downard, 1997). Fibrils play an important role in development and vegetative growth. Fibrils are composed of roughly equal amounts of polysaccharides and proteins, forming a “velvety-coat” over the cell. The protein component of the fibrils has been shown to contain a species known as FA-1 (Behmlander and Dworkin, 1991). But the polysaccharide component of the fibril is complex and not well characterized in *M. xanthus*. Bacterial polysaccharides play a variety of roles related to attachment, biofilm formation and motility. In both *Pseudomonas* and *M. xanthus*, polysaccharides are involved in biofilm formation (Davies *et al.*, 1998; Ramaswamy, 1997). In addition, fibril polysaccharides take an important part in development, tactile sensing (Lee *et al.*, 1995), and generating A-motility forces in *M. xanthus* (Wolgemuth *et al.*, 2002; Kaiser 2003). More importantly, Li and collaborators (Li *et al.*, 2003) show that extracellular polysaccharide appears to have a component that is required for pilus retraction and social motility, which results from it. This component could be replaced by chitin (poly-[1→4]-β-N-acetyl-D-glucosamine) which restored pilus retraction and S-motility.

Myxococcus xanthus produces high amounts of extracellular polysaccharides under normal conditions both in the vegetative cycle and the multicellular developmental cycle. In wild-

type cells, exopolysaccharide content increases significantly as cells enter the stationary phase of growth or upon addition of Ca^{2+} to growing cells, and the polysaccharide-induced cells exhibit an enhanced capacity for cell-cell agglutination (Kim *et al.*, 1999). The basic units of the polysaccharides are galactose, glucosamine, glucose, rhamnose, and xylose (Behmlander and Dworkin, 1994). Interestingly, glucosamine, and to a lesser degree glucose and galactose as well, strongly inhibit cell-cell cohesion which is required for development. On the other hand, Ca^{+} dramatically improves the cohesion in the wildtype.

Most research in *M. xanthus* is focused on its development because it provides a simple system to study many genes and phenomena associated with development in general. The *M. xanthus* developmental process is characterized by a temporal cascade of development-specific gene expression, which is absolutely dependent upon cell-cell signaling. Five classes (A to E) of intercellular signaling mutations have been identified, designated *asg* for A-signal, *bsg* for B-signal, etc., each of which arrests development at a characteristic stage. Each class of signaling mutant is thought to be defective in the production of a distinct class of extracellular signal that is required for continued progress through the developmental program.

M. xanthus development goes through a series of structural changes to form the fruiting bodies and sporulate at the end of the cycle. The earliest step is called traveling waves, which takes place shortly after dense cells are spread on starvation media with a solid surface, such as TPM agar. In traveling waves, cells move back and forth in an effort to find the appropriate cell concentration (quorum sensing) for shifting cellular processes into the development cycle and coordinating the spatial distribution of aggregation centers for

fruiting body formation.

The *M. xanthus* development program is a combination of the structural changes and a spatially and temporally regulated signaling system by which the appropriate signaling pathways are turned on and off at the right time in very narrowly localized places (at submillimeter scale). For example, at about 2 hours after cells are spread on development media, the A-signal is turned on. A-signal is a complex mixture. A part of the A-signal mixture is composed of amino acids (in μM range), such as tyrosine, proline, phenylalanine, tryptophan, leucine, and isoleucine. The rest of the A-signal mixture is not identified. A detailed A-signal study is presented in Chapter 3, where evidence is shown that in addition to amino acids, short chain fatty acids and pyruvate have A-factor activity. However, at present time, it is not clear whether these non-amino acyl compounds are part of the A-signal mixture. A-signal is a quorum-sensing signal that appears to be released into the media by the starving cells. At an appropriate concentration (in μM range), it triggers the cells to perform developmental functions. At a lower or higher concentration, the A-signal may dramatically change the behavior of the culture. For example, at a micromolar level, amino acids trigger development, but at a millimolar level they promote vegetative growth.

The B-signal is turned on immediately after cells are subjected to the developmental conditions, and persists till 20 hours into the development process. The gene *bsgA* was found to be identical with another ATP-dependent protease gene *lonD* found in *M. xanthus* (Tojo *et al.*, 1993). The protein LonD is homologous to the ATP-dependent protease Lon from *Escherichia coli*, which is involved in stringent response. The B-signal is believed to

be the product of an intracellular ATP-dependent protease, BsgA (Gill *et al.*, 1993; Tojo *et al.*, 1993). Its chemical nature is not clearly known yet. Nevertheless, it is thought to be some molecule(s) produced due to the intracellular proteolysis by BsgA (Hager *et al.*, 2001).

The third developmental signal, C-signal, is turned on at around 3 hours after the onset of the development process and lasts to ~23 hours (Sogaard-Andersen and Kaiser, 1996; Lobedanz and Sogaard-Andersen, 2003). The C-signal is coded by the gene *csgA*. The *csgA* gene encodes two proteins however. One is the full-length protein at ~25 kD. The other, at ~17 kD, is synthesized by N-terminal proteolytic processing of the full-length protein. In vegetative growth, protein CsgA is a full-length 25 kD cell-surface, short chain alcohol dehydrogenase, the 17 kD version does not exist. But once the cells enter the development program, a serine protease gradually converts the 25 kD CsgA into the 17 kD carboxyl terminal fragment, which plays the role of the C-signal (Sogaard-Andersen and Kaiser, 1996; Lobedanz and Sogaard-Andersen, 2003; Kim and Kaiser, 1991). Both the 25 kD and 17 kD versions of CsgA are cell-surface bound. C-signaling requires cell-cell contact through the tactile sensory system (Lee *et al.* 1995). Different concentrations of C-signal controls the timing of developmental gene expression and spacing of the aggregation centers via a process called traveling waves in which cells reverse traveling directions on contact with the cells moving in the opposite direction (Lobedanz and Sogaard-Andersen, 2003; Kim and Kaiser, 1991).

The fourth signal, D-signal, has not been identified. The *dsg* gene encodes a translation initiation factor, a homolog of IF3 of *Escherichia coli* (Cheng and Kaiser, 1994; Kalman

and Kaiser, 1994). Understandably, *dsg* is essential for growth. Generally, cells with defects in the D-signaling system can still aggregate and sporulate. But aggregation in these strains leads to abnormal structures and is significantly delayed, and the rate of sporulation is dramatically reduced. The D-signal seems to have multiple functional components because one mutant of the *dsg* class grows well vegetatively but fails to form fruiting body completely. It is likely that the actual D-signal molecule(s) that effects this signaling pathway could be a product(s) of the Dsg protein directly controlled protein or further downstream. The characteristics of the D-signal molecule(s) is not known at present time.

E-signal is believed to be branched chain fatty acids, or their derivatives. The *esg* locus encodes two components (E1 and E2) of the branched chain α -keto dehydrogenase complex that is required for branched chain fatty acid synthesis and development. The *esg* mutants grown in a medium supplemented with branched chain fatty acids can develop almost as well as the wildtype (Toal *et al.*, 1995). Fatty acids have long been known as signal mediators in mammals. The fatty acid signal in *M. xanthus* is much simpler than the fatty acid signal known in mammals such as prostaglandins (Ferreira and Vane, 1967; Nomura *et al.*, 2004). The E-signal appears at the 6th hour in the development process. In wild type, branched chain fatty acids are incorporated into the outer membrane during vegetative growth, and released during development, probably by phospholipase, to act as the E-signal controlling the downstream processes in the development program (Bartholomeusz, 1998).

All these developmental steps were elucidated in polysaccharide-producing strains of *M. xanthus*, the commonly used strain DK1622. These strains all display a rough colony morphology. One exception however is the strain FB and its derivatives. FB carries three

point mutations in the gene *pilQ* (thus *pilQ1*, Wall *et al.*, 1999), has a smooth colony morphology, and produces much reduced levels of polysaccharides which is easily detected using Calcofluor White supplemented agar plates (Ramaswamy *et al.*, 1997), as do other smooth strains. However, FB carries out the complete normal development process. In this work, a large number of smooth strains (each carries an insertion mutation in various genes) exhibit defects in development. The contrast between FB and other smooth-looking strains brings up the question, what genes are involved in control and synthesis of fibril polysaccharide and what is the relationship of these genes to development.

For these reasons, we generated a large number (~10,000) of transposon insertion mutants for selecting smooth-looking, and Calcofluor White-binding deficient strains (>80), sequenced the insertion points (26) from the transposon termini, developed computer programs to stream line the insertion point mapping processes, and designed internal fragment replacement mutagenesis protocol to efficiently analyze the genes around the insertion sites.

MATERIALS AND METHODS

Bacterial Strains And Culture Conditions.

The wildtype *Myxococcus xanthus* strain DK1622 and FB were from Kaiser's laboratory. The mutant strains SR53, SR171, SR200, and *esg* were documented previously (Ramaswamy, 1997). All cds series strains were generated in this work as described in the Transposon Mutagenesis method below. Unless indicated otherwise, the cultures of *Myxococcus xanthus* strains were grown at 30°C, in Casitone Yeast Extract (CYE) broth (Casitone 10g/L, Yeast Extract 5g/L, MgSO₄·7H₂O 1g/L) in a flask shaking at 250 RPM or plated on CYE agar (0.15% w/v). Roller drum grown cells were inoculated in a Φ15x150 mm glass tube on a New Brunswick TC-7 roller drum rolling at 50 RPM. Log phase cells needed for experimentation were collected between 80-100 Klett units (100 Klett units is equivalent to ~0.7 Absorbance, and approximately 5 to 7 x 10⁸ cells/ml). *Escherichia coli* DH5α grown in LB medium is used as the cloning host, and plasmid source.

Electrocompetent Cell Preparation.

Electrocompetent *E. coli* and *M. xanthus* cells were prepared by essentially the same procedure. Mid-log phase cells were harvested by centrifugation at 6000 rpm for 10 min. The pellet was washed twice by resuspending the cells in one volume of deionized and distilled water (*M. xanthus*) or of 10% glycerol (*E. coli*) and re-pelleting. Cells are prepared just before use, although *E. coli* competent cells can be deep frozen for later use.

Plasmid Isolation

Regular plasmid isolation was performed according to the procedures outlined in the

Molecular Cloning: A Laboratory Manual by Sambrook *et al.* (1989). High purity plasmid DNA for sequencing was obtained using Qiagen's Plasmid Midi Kit (Qiagen Inc., Valencia, Canada) as described by the manufacturer.

Transposon Mutagenesis

M. xanthus (DK1622) was grown to exponential phase and collected by centrifuging at 8000 x g for 10 min. Pellets were washed twice in deionized and distilled water. Final pellets were resuspended in ~1/50 volume of deionized and distilled water. Then 40-100 μ l concentrated cells were mixed with 0.01-0.1 μ g plasmid DNA carrying the transposon *magellan-4* (derived from *mariner*, and carried on the plasmid pMycoMar) (Rubin *et al.*, 1999; Youderian *et al.*, 2003) (Figure 1.1), and immediately subjected to an electric shock at 800 volts for 5-8 milliseconds. Cells were resuspended in CYE immediately and allowed to recover for about four hours at 30°C with shaking. Eventually about one tenth of the recovered cells were plated on a Calcofluor White containing (50 μ g/ml) CYE plate with kanamycin at 40 μ g/ml. After 4-7 days of incubation, colonies that were smooth looking and defective in producing fluorescence were picked for further analysis.

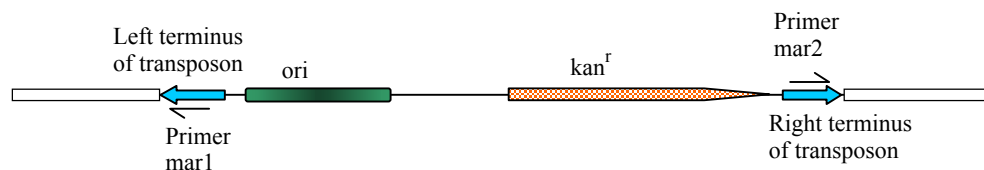


Figure 1.1. The transposon *magellan-4* structure as it is inserted in the genome.

Design of internal fragment replacement mutagenesis

The design of internal fragment replacement mutagenesis utilizes the property of PCR to

target a specific site and the homologous recombination that takes place naturally in *M. xanthus*. An internal fragment of a target gene is amplified using PCR, and cloned into the vector pZerO-2 (Invitrogen, Carlsbad, California). This vector with the cloned target fragment is electroporated into wildtype *M. xanthus* (DK1622) cells. Since the plasmid vector pZerO-2 does not replicate in *M. xanthus*, kanamycin resistance could be maintained only via integration of the plasmid into the host genome by homologous recombination between the cloned fragment on the vector and the original gene on the genome (Figure 1.2). The recombinants are easily picked up by plating the cells on an appropriate antibiotic (kanamycin) containing plate and selecting for antibiotic resistance clones (Figure 1.2). For this design to work, the PCR amplified fragment has to be an internal fragment of the gene so that after recombination the functional gene is not regenerated, neither upstream nor down stream from the recombination site. PCR primer sets (Table 1.1) were designed to produce an internal fragment for each of the seven genes (glycosyltransferases and genes sharing operons with them, see Figure 1.11 for more information). An EcoRI restriction enzyme site was added to the 5' ends of the forward and reverse primers. An extra "GC" was added to the 5' of the EcoRI site. Primers were chosen based on the desired length (21-27 base pairs), the GC content, and melting temperature (74°C-76°C using the formula by R. Owczarzy, 1998). Primers were ordered from Integrated DNA Technologies, Coralville IA.

Genomic DNA isolation.

The genomic DNA from *M. xanthus* was isolated using the DNeasy Tissue Kit from Qiagen (Valencia, Canada). Cells were grown to mid-log phase, and approximately 10^9 cells were

collected by centrifugation. The cell pellet was stored frozen at -70°C until the DNA isolation procedure was performed.

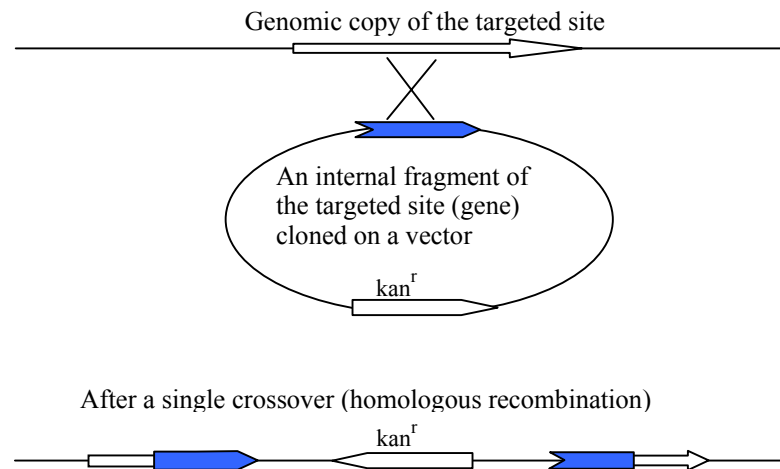


Figure 1.2 Schematic drawing to show how the Internal fragment replacement mutagenesis works.

Cloning

Since the magellan-4 transposon we used carries an *ori* site and a kanamycin resistance gene, it is self-replicative in *E. coli*. Genomic DNA of the transposon mutants was digested with restriction enzyme SacII, which does not cut inside the transposon. The digested genomic DNA was ligated with T4 ligase for one hour, then mixed with electrocompetent *E. coli* cells (DH5 α) and subjected to an electric shock at 1500-2500 volts for 3-8 milliseconds. Electroporated cells were immediately resuspended in LB broth and allowed to recover for one hour. One tenth of the recovered cells were plated on a kanamycin containing LB plate. After one day incubation, the colonies appeared on the plate were picked and grown for plasmid isolation. If the plasmid yielded the expected banding patterns on electrophoresis gel the colonies were collected for further analysis.

Table 1.1 Primers used for internal fragment replacement mutagenesis

Predicted gene	ORF	Primer Sequence	Position (end base)	Result frag. (bp)
GumB	29_2	For 5'-CGGAATTC CCCGGGGCCTGGGCAAGTACAC -3'	3848->	445
		Rev 5'-CGGAATTC CCGCATGACGAAGATGCGGTCCTT -3'	<-3420	
?* high GC	29_3	For 5'-CGGAATTC TGTCCCTGCGCCTGTCCGAGG -3'	1266->	885
		Rev 5'-CGGAATTC CCTCCACGTCCCGCGGAGGTCG -3'	<-2134	
GumC	29_5	For 5'-CGGAATTC GCAGGCGAAGCTGGTGGAGTAC -3'	2825->	1336
		Rev 5'-CGGAATTC CGGACTCCAGCTCACCCCGGGTC -3'	<-4175	
GT**	29_6	For 5'-CGGAATTC TCCTGGAGCGGCAGGACACGA -3'	4675->	1162
		Rev 5'-CGGAATTC CATGGCATCGGGTTGTGGGACC -3'	<-5321	
WecG	53_3	For 5'-CGGAATTC CGCCAACGTGGACCACGTGG -3'	<-2010	503
		Rev 5'-CGGAATTC CGTCCCGCGATGAAGTCCAA -3'	1524->	
CelA	53_4	For 5'-CGGAATTC GGTGCTGCGCGCTCCAAGAG -3'	<-3828	1131
		Rev 5'-CGGAATTC GGGCTCGCCAATCCAGGACTCG -3'	2714->	
GT	53_6	For 5'-CGGAATTC GGCCCTGGCCCTGCCCTACAC -3'	<-5876	587
		Rev 5'-CGGAATTC GTACGCACGTCCGGGTGCTT -3'	5306->	

* The identity of the gene is not known. ** GT = glycosyltransferase.

Social Motility Assay

Exponentially growing cells were pelleted and resuspended in CYE ($\sim 5 \times 10^9$ cells/ml), then spotted on the surface of CYE soft agar (0.3% w/v) and incubated for 72 hours. The diameter of the spots (including the fringe) minus the original spot size (6 mm) is the net expanding movement the cells made during incubation. The ratio of the net expanding movement of a mutant to that of the wildtype was taken as the relative motility of the mutant.

Agglutination (Cohesion) Assay

Agglutination buffer (10 mM MOPS [pH6.8], 1mM MgCl₂, 1 mM CaCl₂), and MOPS buffer were made according to the procedure outlined by Shimkets *et al.* (1986). The wildtype as well as mutant strains were grown in CYE broth to 60-100 Klett units. 15 ml of each culture were collected by centrifugation at 8000 g for 5 min. The pellet was resuspended in MOPS, at about 50 x the initial concentration. This suspension was ready for addition to the agglutination assay buffer for characterization. The agglutination was measured in a Klett-Summerson meter with a red filter (maximum passing wavelength 620 nm), or in Spec 20 with wavelength set at 600 nm. The turbidity of the suspension was taken at an interval of 5 min or as indicated otherwise. The ratio of the turbidity at a given time to the initial turbidity was plotted against time, resulting in a cohesion (agglutination) curve.

Calcofluor White-Binding Assay

Previous experiments show that polysaccharide-producing cells bind Calcofluor White and become fluorescent under UV light. Mid-log cells were concentrated to ~1500 Klett units and 10 µl each was spotted on CYE plates containing 0.3% agar and 50 µg/ml Calcofluor White (Sigma, St. Louis, Missouri). Plates were incubated at 30°C for 40 hours or longer. Colonies were observed with a hand held UV light source (366 nm) to determine the level of the spots' fluorescence. Representative photos were taken with a Nikon Coolpix digital camera under a combination of ultraviolet light and incandescent (tungsten lamp) visible

light, or with a Cannon Power G2 digital camera in absolute darkness under a single handheld ultraviolet light source. The relative brightness of fluorescence is analyzed visually based on the digital photographs.

Multicellular Development assay

Two methods of development assay were employed. (1) Mid-log cells were concentrated by centrifugation. 10 µl of concentrated cells (~1500 KU) were spotted on TPM (10 mM Tris [pH7.5], 1 mM KH₂PO₄, 8 mM MgSO₄) agar (0.15% w/v, high purity) plates, and incubated at 30°C for ~48 hours. Multicellular structures were examined under a microscope to determine the deficiencies in development for each mutant strain. Photographs were taken as above.

The alternative method of development assay: (2) Mid-log cells were concentrated by centrifugation. 10 µl of concentrated cells (~1500 KU) were spotted on CF (10 mM MOPS[pH7.6], 0.015% Casitone, 8 mM MgSO₄, 1 mM KH₂PO₄, 2% Na Citrate, 1% pyruvate) agar (0.15% w/v, high purity) plates, and incubated at 30°C for ~96 hours. Multicellular structures were examined under a microscope to determine the deficiency in development for each mutant strain. Photographs were taken through a microscope with a very low magnification objective lens.

DNA Sequencing

Primers for sequencing insertion mutants are: mar1 5'-CGCCATCTATGTGTCAGACCGG

GG-3', and mar2 5'-TGTGTTTTTCTTTGTTAGACCG-3', which respectively anneal to the left-end and the right-end of the magellan-4 transposon extending outward. High purity plasmid DNA was prepared as described above. The DNA samples together with the primers were sent to the Oklahoma Memorial Research Foundation facility for sequencing. All sequences were sequenced only once, with undetermined bases marked as N. No error checking or correction was done on the sequencing readout, and it was directly used for insertion point mapping.

Sequence Analysis Strategies

There are a large number of sequence analysis software tools available with paid or free access. However, the combination of the Artemis program from The Sanger Institute (Rutherford *et al.*, 2000), BLAST from GenBank (Altschul *et al.*, 1990), and our own specialized program SHAPE (see Chapter 2) worked best for our data. Our sequence analysis was based on the sequences retrieved from the NCBI *M. xanthus* database. Therefore, the sequence reliability is much higher than our “single pass” sequences used for mapping the insertion points. Since *M. xanthus* genomic DNA has a high G+C ratio, the raw sequences were evaluated and annotated first with the sequence analysis (computer program) tool Artemis for G+C biases on the three positions of each codon, and potential open reading frames (ORF). Then the raw DNA was divided into trunks or ORFs, or translated into amino acid sequences (all within the Artemis program) before being subjected to the BLAST search via NCBI webpages. This step was to make the query sequence short (e.g. single gene length), which was necessary because we have very long sequences (contigs) in the analysis, and was essential for taking the full advantage of the

BLAST program's search sensitivity and efficiency. It is always possible to have errors in the query sequences, an extra base may set off the codon reading frame and miss important matches to the database sequences. Therefore, BLASTX (Altschul et al., 1990) searches were used to establish basic similarity relationships between our query sequences and what has been accumulated in the GenBank databases at that moment in time. Within the time frame of this project, GenBank DNA databases have changed significantly from the viewpoint of the query results for our sequences.

The sequences flanking the insertion points were obtained as described in the method for DNA Sequencing. The mar1 sequences (mar1 primer anneals to the upstream end of the *magellan-4* transposon, Figure 1.1) are the chromosomal sequence fragments flanking the mar1 end of the insertions. The mar2 sequences (mar2 primer anneals to the downstream end of the transposon) are the genomic sequence fragments flanking the mar2 end of the insertions. The sequencing was done only once for each insertion clone. Therefore, the sequence data we obtained is not expected to be of very high quality because lack of redundancy leads to inaccuracy. The sequences flanking the insertions were used to search sequence databases. Initially, the proprietary *M. xanthus* contig database at Cereon, a subsidiary of Monsanto was used. A condition on the use of this database was to retrieve only limited sequence from the contig corresponding to our query sequence. The sequence data were also used to query the GenBank for previously published sequences. In 2004, it became possible to search and retrieve sequences from the databases of incomplete genomes (including *M. xanthus*) at the GenBank.

Then retrieved contigs (Rcontigs) were built with the DNA sequence fragments from the Cereon *M. xanthus* database, the GenBank databases, and our own insertion point sequences. Many insertion sequences (>40) from the Calcofluor White-binding deficient mutants were mapped by means of a specialized computer program (called SHAPE, see chapter 2). A mini database of the retrieved contigs was established to permit management of the information on chromosomal sequence and in particular the flanking regions of the insertions. The actual mapping begins with running the SHAPE program on our database. The program takes the insertion sequences as the query sequence to scan every retrieved contig to find matches. The search algorithm is superficially similar to BLAST searches. This process generates a data set for each and every possible matching pair between insertion sequences and the Rcontigs. The mapping results can be viewed via a web interface (e.g. a web browser). The outcome is the precise base position of the transposon insertion. The details of the computer programs are explained in Chapter 2.

As a part of this informatics treatment of *M. xanthus* sequence data, it was realized that the eight large contigs available could be assembled with the help of the physical map (He *et al.*, 1994), albeit with gaps between the contigs. A small computer program was developed to find the restriction sites in the eight contigs and compared with the restriction fragments on the physical map. Then, the whole genome sequence map is reconstructed in a way very similar to the way a plasmid map is reconstructed from restriction digestions. This assembled contig (sequence) map is called the gapped genomic sequence map (Figure 1.10).

To determine an ORF in *M. xanthus* sequence is slightly easier than in other organisms

because the high G+C bias of the genome lends help in ORF determination. In high G+C organisms, the authentic ORFs usually carry a high G+C bias on the third base of the codons. Therefore in judging the authenticity of various ORFs, one can relatively easily compare which ORF carries a high G+C bias on the third base of the codon. This function is provided in the computer program Artemis, called “GC frame plot”. This genomic feature enabled us to predict the hypothetical proteins to a better certainty even without a database match. In addition, the hypothetical ORF sequences can be used to search the GenBank database to find out whether there are homologues known in the database.

RESULTS AND DISCUSSION

Growth Characteristics *M. xanthus* Development Mutants

After electroporations of several *M. xanthus* DK1622 (wildtype) cultures with transposon *magellan-4* carrying plasmid (pMycoMar), more than 10,000 kanamycin resistant (kan^r) strains were obtained. Of these mutants, more than 80 were Calcofluor White-binding deficient; 43 of these were studied further as developmental mutants. Characterization of each strain included motility, cohesion, Calcofluor White-binding, growth curves under different aeration conditions, development, and insertion site sequence analysis.

Observations of roller-drum grown cultures of these strains suggested that some of them formed clumps and precipitated more rapidly than others when agitation was stopped. The phenomenon was much less obvious in shaking flasks. It is not known if the different patterns were due to aeration differences, shear, or some other factor, but as will be described below, these were correlated with smooth/rough colony morphology, cohesion, Calcofluor White-binding, and development.

Growth curves for each of the insertion mutants and a few other strains in CYE medium are displayed in Figure 1.3 (pages 20-23). Two growth conditions are shown: pink growth curves represent roller-drum growth; blue curves, shaker flask growth. Panels 1 and 2 (page 20) are the wildtype rough strain DK1622 and smooth-looking strain FB, respectively. The smooth strain did not clump under roller-drum growth (pink and blue growth curves were superimposed); whereas the roller rough wildtype strain appeared to lag for several hours and after maximum absorbance was reached the culture density decreased quickly.

The panels followed *cds1* through *cds42* are the results with the insertion strains. In those panels with a pink growth curve, the pattern was somewhat similar to the wildtype, when a blue curve only is shown, the roller drum growth did not differ from shaker flask culture. The final five panels demonstrate previously characterized developmental mutants. The final four mutants (namely SR171, SR200, SR483 and *esg*) are in loci unrelated to the insertion mutants studied here.

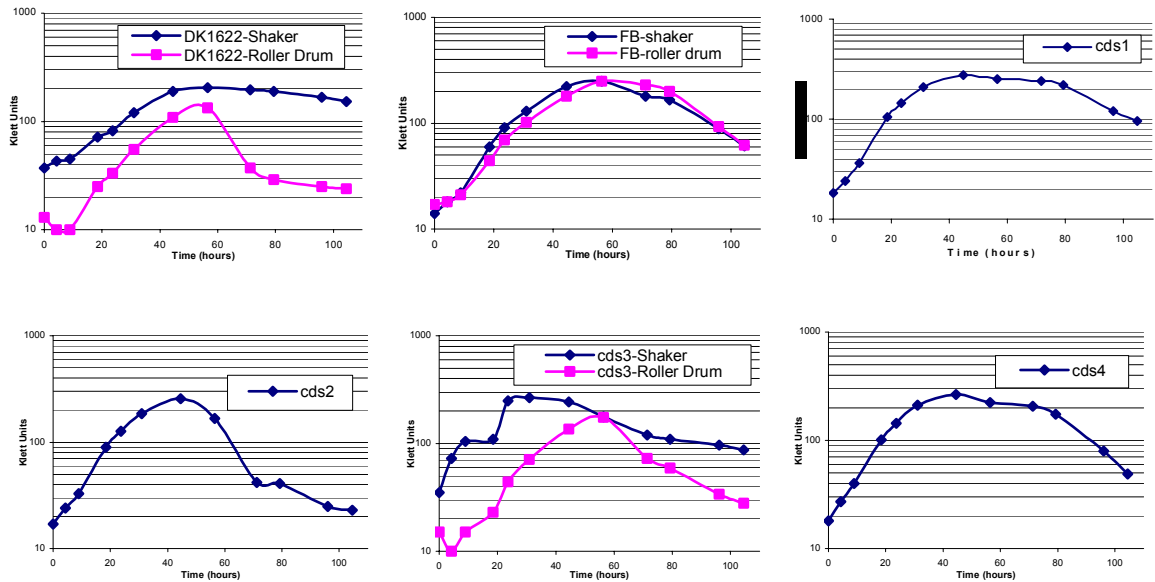


Figure 1.3 Growth curves of *M. xanthus* strains in CYE at 30°C. (continued on next page)

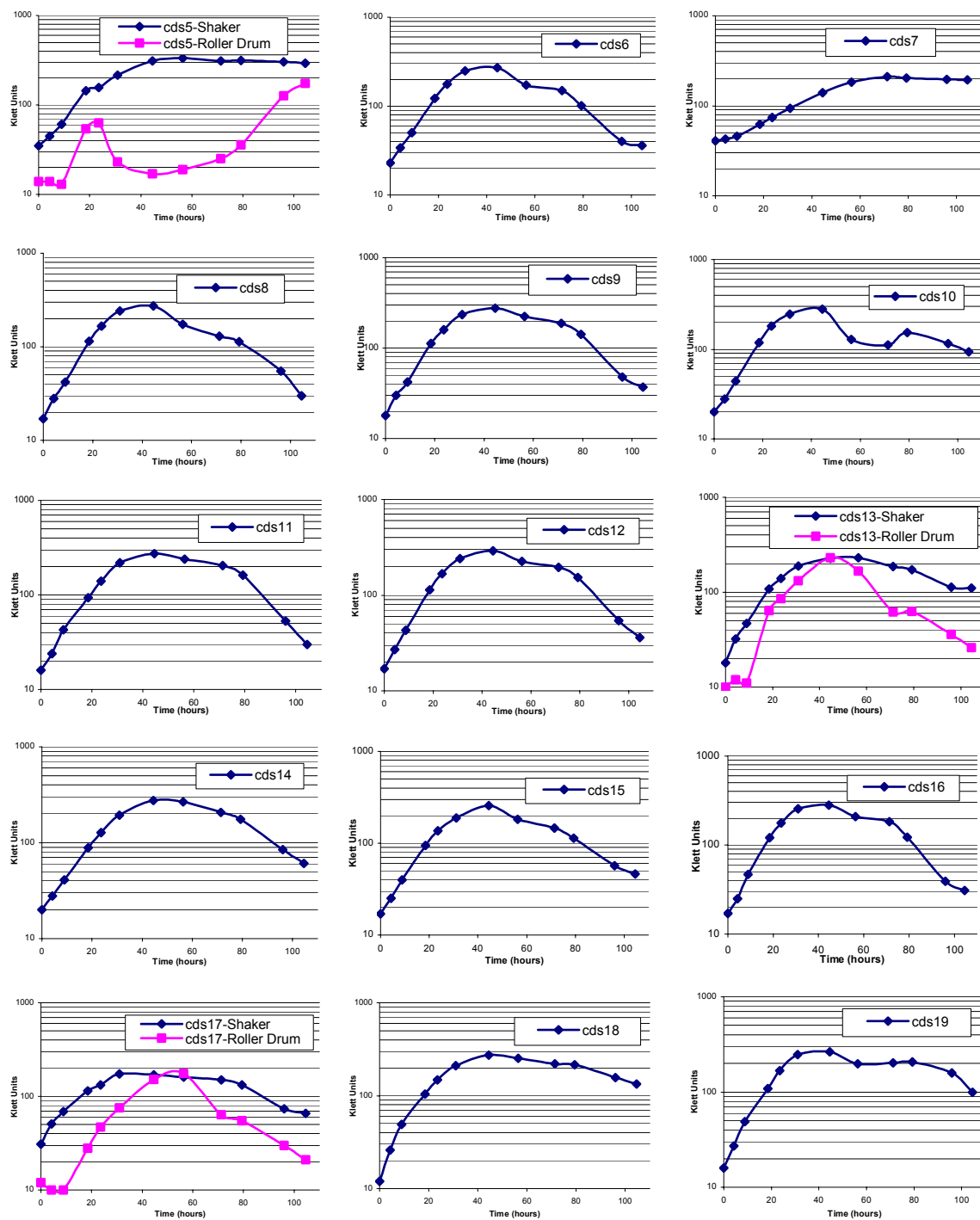


Figure 1.3 Growth curves of *M. xanthus* strains in CYE at 30°C. (continued on next page)

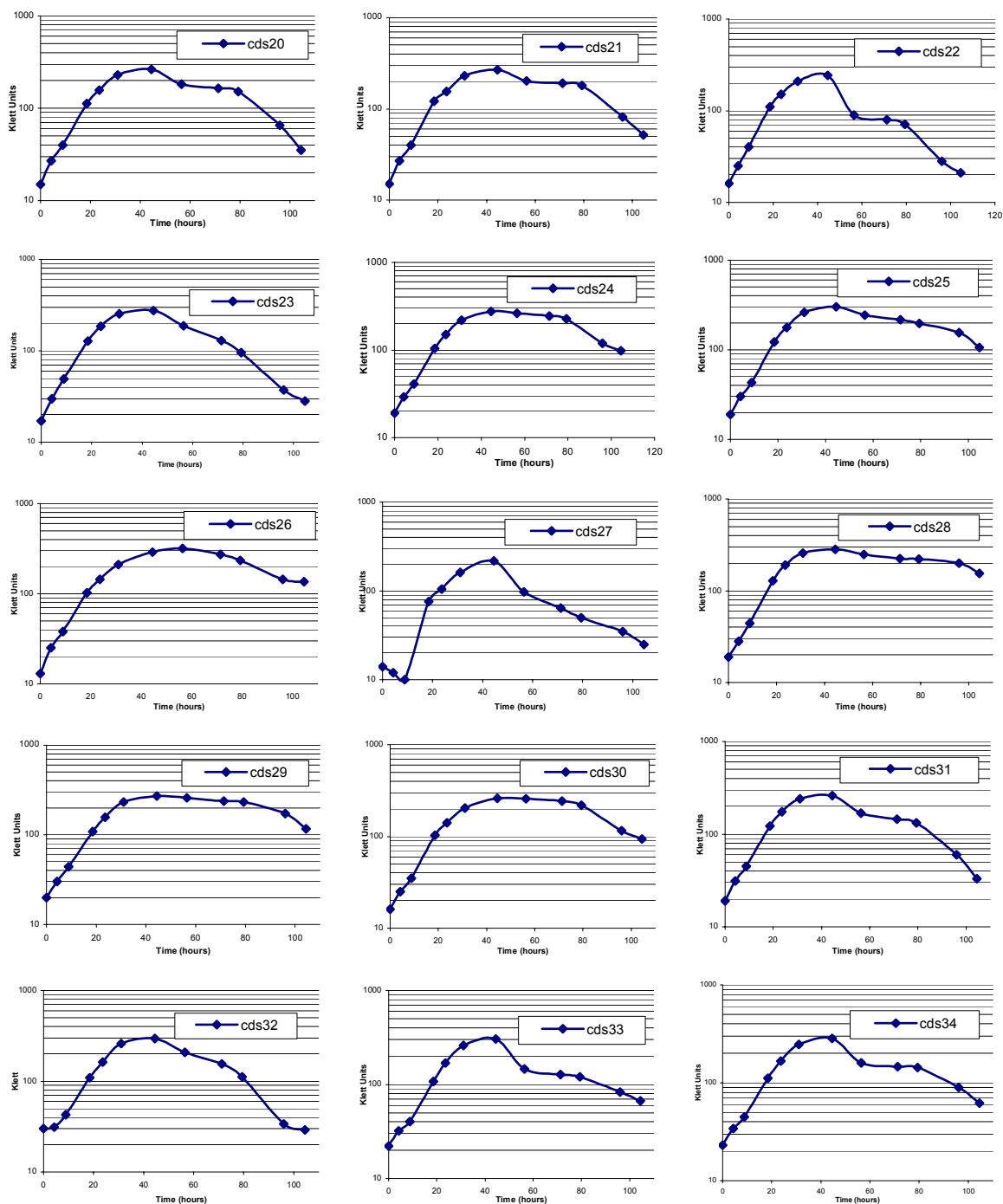


Figure 1.3 Growth curves of *M. xanthus* strains in CYE at 30°C. (continued on next page)

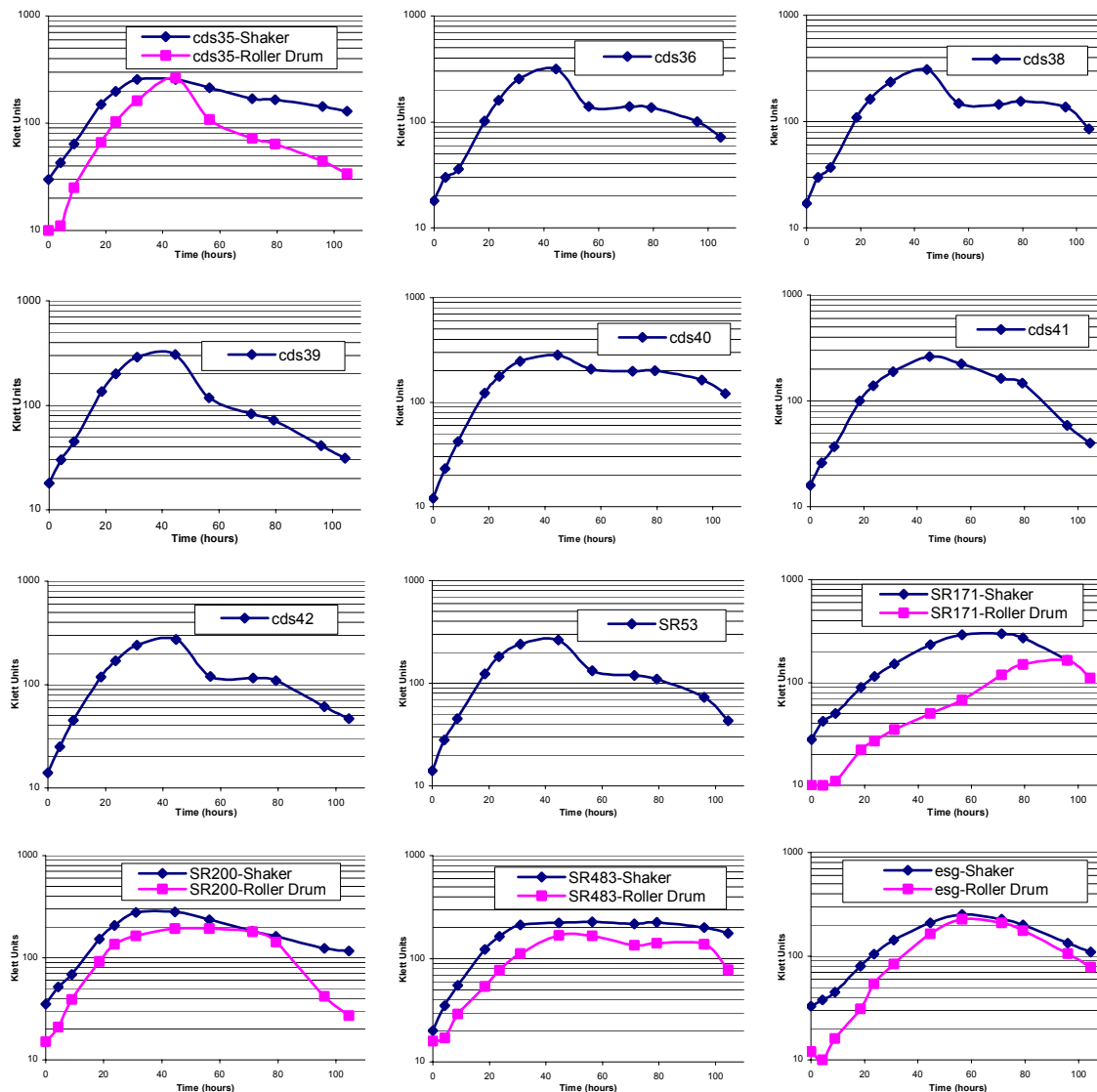


Figure 1.3 Growth curves of *M. xanthus* strains in CYE at 30°C. The wild type strain DK1622 (panel 1, page 20) is used as a reference. Other previously characterized strains (FB [panel 2, page 20], SR53, SR171, SR200, SR483, and esg [last five panels, page 23]) are also included to compare the variability in the growth curve between shaker grown (blue) and roller-drum grown cultures (pink).

The simple use of two growth regiments: roller-drum (with slow rotation [50 RPM], low shear); and shaker flask (rapid rotation [250RPM], higher shear permitted strong distinctions to be drawn between some strains. The more complex assays (e.g.

agglutination) that follow yield results that seem to correlate with the growth curve pattern. Understanding these growth patterns is critical to permit one to obtain cultures of each strain that are equivalent in cell density and stage of growth for more quantitative analyses of developmentally related traits.

When the growth curves were studied in strains with other phenotypes, a trend of inverse correlation is obtained between the maximum growth concentration (the highest concentration a culture can reach, measured in Klett Units) and the cohesion efficiency, and between the maximum growth concentration (Klett Units) and the motility of the strain. This observation coincides with a variant of the cohesion assay (also called clumping assay, Wu *et al.*, 1997) where a grown cell suspension is left standing in a tube and its turbidity is monitored. The growth curves are shown in Figure 1.3, where a variety of growth patterns was observed.

It was observed that the different mutants grew quite differently even when incubated in the same medium, and same container, with shaking at the same speed. Some strains (eg. *cds2*, *cds8*, *cds10*, *cds33*) maintained a very short period of viability once they reached the maximum concentration of cells, as indicated by the exponential decrease in absorbance. Some strains (*cds5* and *cds7*) grew very slowly. For example, *cds7* shows a doubling time well over 10 hours, compared to strain DK1622 (wildtype) which doubled in less than 5 hours. The difference in growth pattern between the roller drum tube-grown and shaker flask-grown cultures is particularly noticeable in cohesion proficient strains such as DK1622, *cds3*, *cds5*, and *cds17*). Generally speaking, the shaker flask-grown cultures maintained viability for a longer period in the stationary phase. Another interesting

observation is that two cds strains cds5 and cds7 maintained the cell concentration at or near the maximum for a very very long time (over 60 hours) in the stationary phase (panels 1 and 3, p. 21). However, these two strains have not been cloned, nor sequenced. The genes involved in this phenomenon are unknown.

Cohesion Assay

As was suggested in the previous set of experiments, the various developmental strains exhibit different aggregation behaviour. The cohesion assay measures turbidity changes over time, and reported here as relative turbidity which is the ratio of turbidity at time t over time t_0 when a culture is suspended in the cohesion buffer and permitted to stand without agitation. The cohesion curves for the strains under study are shown in Figure 1.4. Those strains whose relative turbidity decreased to below 10% within three hours of cohesion are considered cohesion-proficient (such as *cds3*, *cds13* and the wildtype DK1622). Cohesion-deficient strains maintained the relative turbidity above 50% after 4 hours (such as strains FB, *cds1*, and *cds8*).

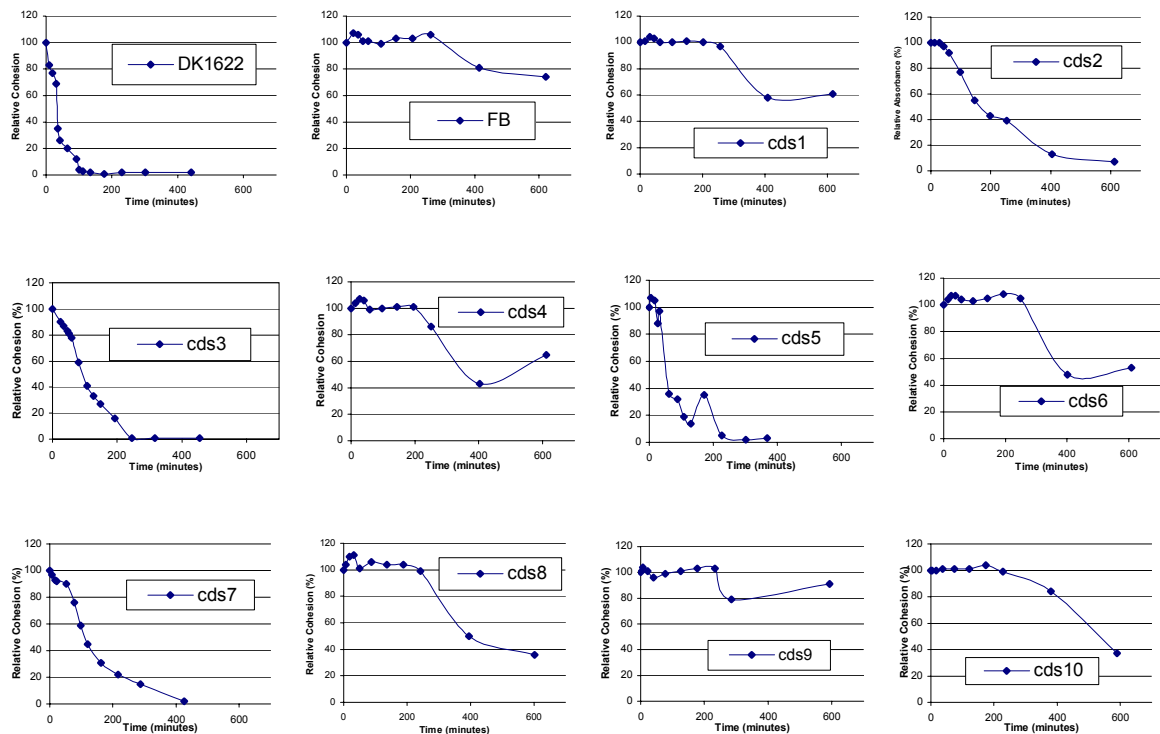


Figure 1.4 Cohesion curves. (continued on next page)

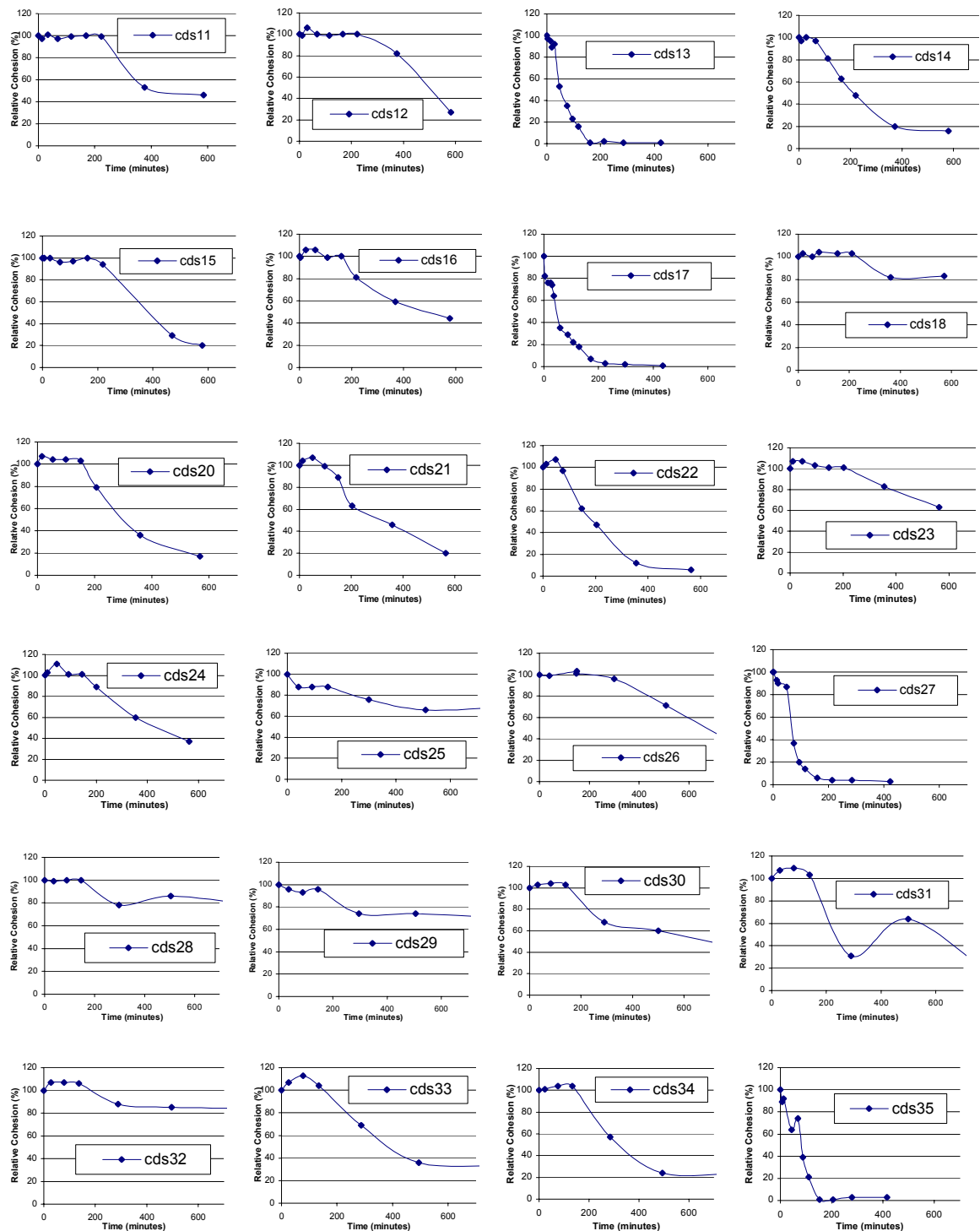


Figure 1.4 Cohesion curves. (continued on next page)

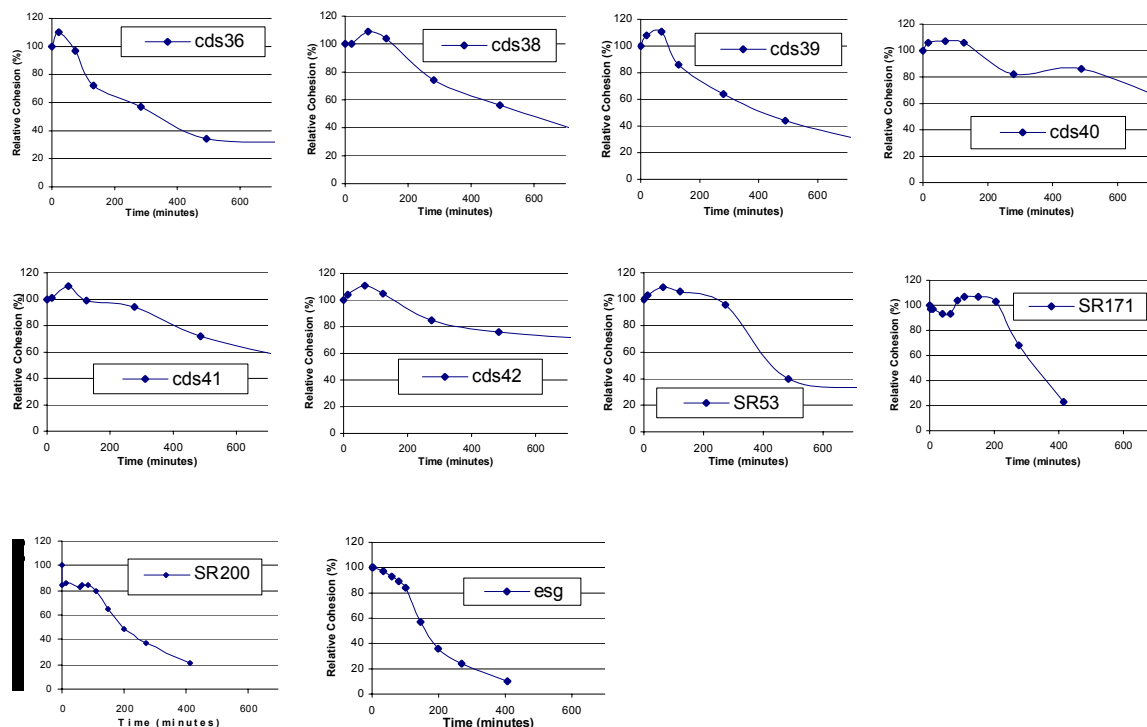


Figure 1.4 Cohesion curves are plotted as the relative turbidity (the turbidity at time t over time t_0) versus time. Strains were assayed for their ability to clump in the cohesion buffer as specified in the Material and Methods section.

The cohesion-proficient strains (DK1622, *cds3* and *cds13*) whose turbidity dropped below 10% of starting turbidity in 3 hours of cohesion assay) were also the strains that did not grow well in roller-drum culture (Figure 1.3) in that they did not reach 250 Klett unit, and then the culture density decreased rapidly. Combined with the Figure 1.3, cohesion proficient strains tend not to grow to very high concentration. For example, DK1622, *cds13*, *cds17*, *cds27*, and *cds35* never reached 250 KU. They all dropped to below 10% of starting turbidity within three hours of the cohesion assay. The results of cohesion assays were identical strains grown in two different rich media CTT and CYE.

Social Motility Assay of Developmental Mutants

Concentrated cells ($\sim 5 \times 10^9$ cells/ml) were spotted on soft agar (0.3% w/v) and incubated for 72 hours. The diameter of the spots (including the fringe) minus the original spot size (6 mm) is the net expanding movement the cells made during incubation. The ratio of the net expanding movement of a mutant to that of the wildtype is taken as the relative motility. The results of S-motility tests on the complete set of developmental mutants and reference strains used in this work are in Figures 1.5 and table 1.2.

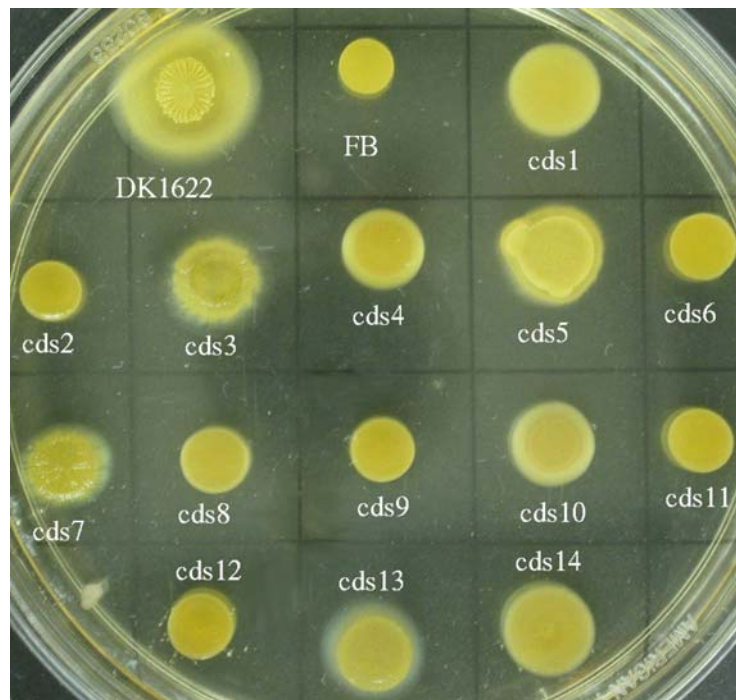


Figure 1.5 Social motility (S-motility) assay of developmental mutants. Cells (5×10^9 ml⁻¹) were plated on 0.3% agar containing CYE in 10 cm plates (nominal), incubated at 30°C for 72 hours. (continued on next page)

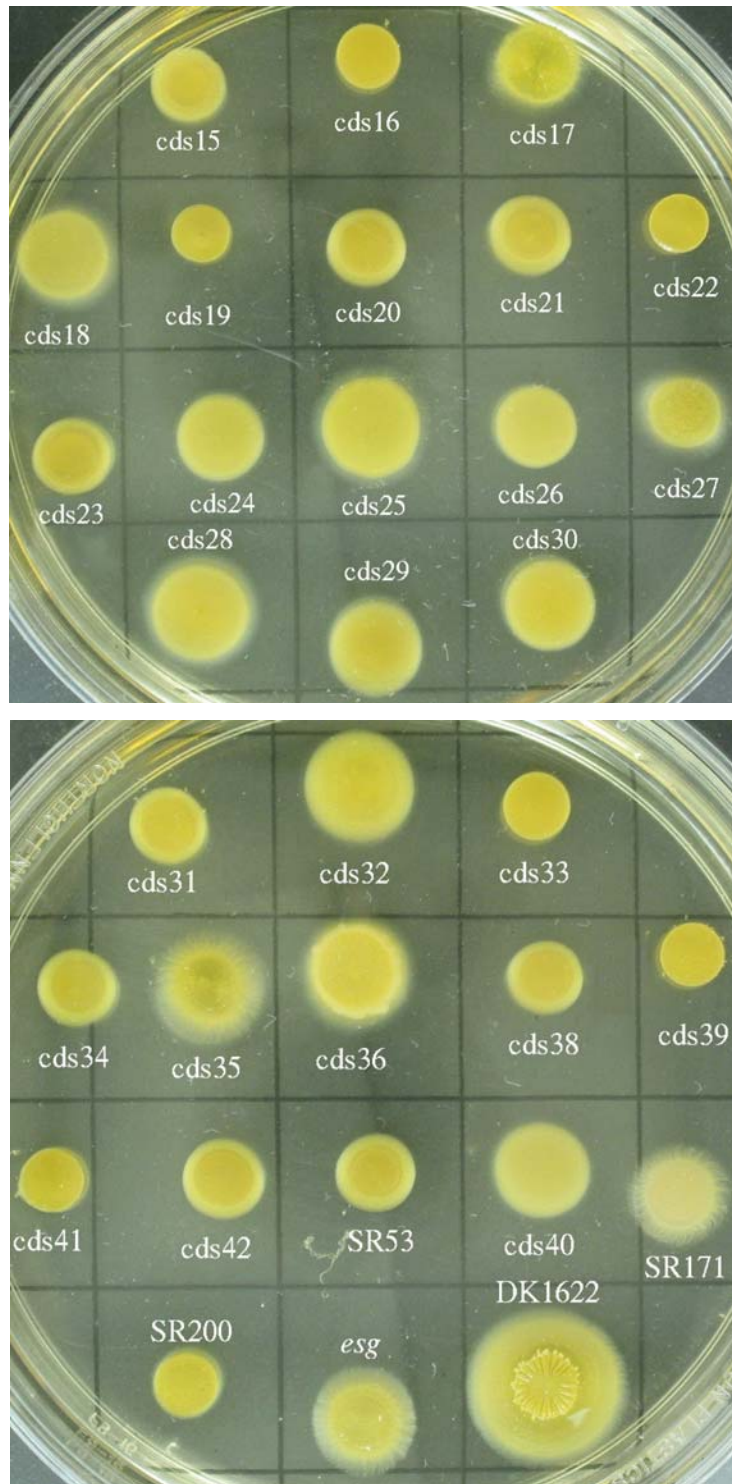


Figure 1.5 Social motility (S-motility) assay of developmental mutants. Cells (5×10^9 ml⁻¹) were plated on 0.3% agar containing CYE in 10 cm plates (nominal), incubated at 30°C for 72 hours. Photographs were taken under normal light with a Canon Power G2 digital camera.

The wildtype DK1622 exhibited the greatest motility of all strains tested. The cohesion-proficient strains tended to retain more of the parent strain's motility. That is, *cds13*, *cds17*, *cds27*, and *cds35* retained more than 60% of wildtype motility (Figure 1.4, Table 1.2). The motility differences among the *cds* strains covered the full range, from S-motility comparable to the wildtype (*cds27*, 77%) to almost completely non-S-motile (*cds22*, 8%).

Table 1.2 Relative social motility

Strain	Diameter (mm)	Expd. Mov. (mm)	Relative S-motility (%)
DK1622	19.0	13.0	100
cds1	12.0	6.0	46
cds2	7.5	1.5	12
cds3	13.7	7.7	59
cds4	9.2	3.2	25
cds5	15.0	9.0	69
cds6	8.3	2.3	18
cds7	10	4.0	31
cds8	8.0	2.0	15
cds9	8.0	2.0	15
cds10	9.6	3.6	28
cds11	8.0	2.0	15
cds12	8.0	2.0	15
cds13	14.0	8.0	62
cds14	14.0	8.0	62
cds15	10.5	4.5	35
cds16	7.7	1.7	13
cds17	14.5	8.5	65
cds18	15.0	9.0	69
cds19	9.0	3.0	23
cds20	11.0	5.0	38
cds21	9.3	3.3	25
cds22	7.0	1.0	8
cds23	8.8	2.8	22
cds24	11.6	5.6	43
cds25	15.0	9.0	69
cds26	10.5	4.5	35
cds27	16.0	10.0	77
cds28	15.0	9.0	69
cds29	13.0	7.0	54
cds30	13.2	7.2	55
cds31	9.6	3.6	28
cds32	14.4	8.4	65
cds33	7.8	1.8	14
cds34	9.5	3.5	27
cds35	15.0	9.0	69
cds36	10.0	4.0	31
cds38	9.2	3.2	25
cds39	8.0	2.0	15
cds40	11.0	5.0	38
cds41	8.3	2.3	18
cds42	10.2	4.2	32
SR53	10.0	4.0	31
SR171	13.2	7.2	55
SR200	10.0	4.0	31
<i>esg</i>	12.7	6.7	52
FB	7.2	1.2	9

Observation of Developmental Defects

This collection of transposon *magellan-4* insertion mutants showed various levels of defects in development. When concentrated cells were spotted on CF agar, the morphology of the cell mass was used to monitor development on CF agar. In Figure 1.6 (the final panel on page 38) the wildtype *M. xanthus* DK1622 is shown to be capable of spreading out from the spot, and forming fruiting bodies within the spot and the adjacent area. The fruiting bodies are round and dark. In contrast, most *cds* mutants (30 out 39) did not form fruiting bodies after 96 hours of incubation (Fig. 1.6). Among the nine strains that formed fruiting bodies, three of them formed fewer than twenty fruiting bodies per spot (Figure 1.6), which is less than 5% of the wildtype level.

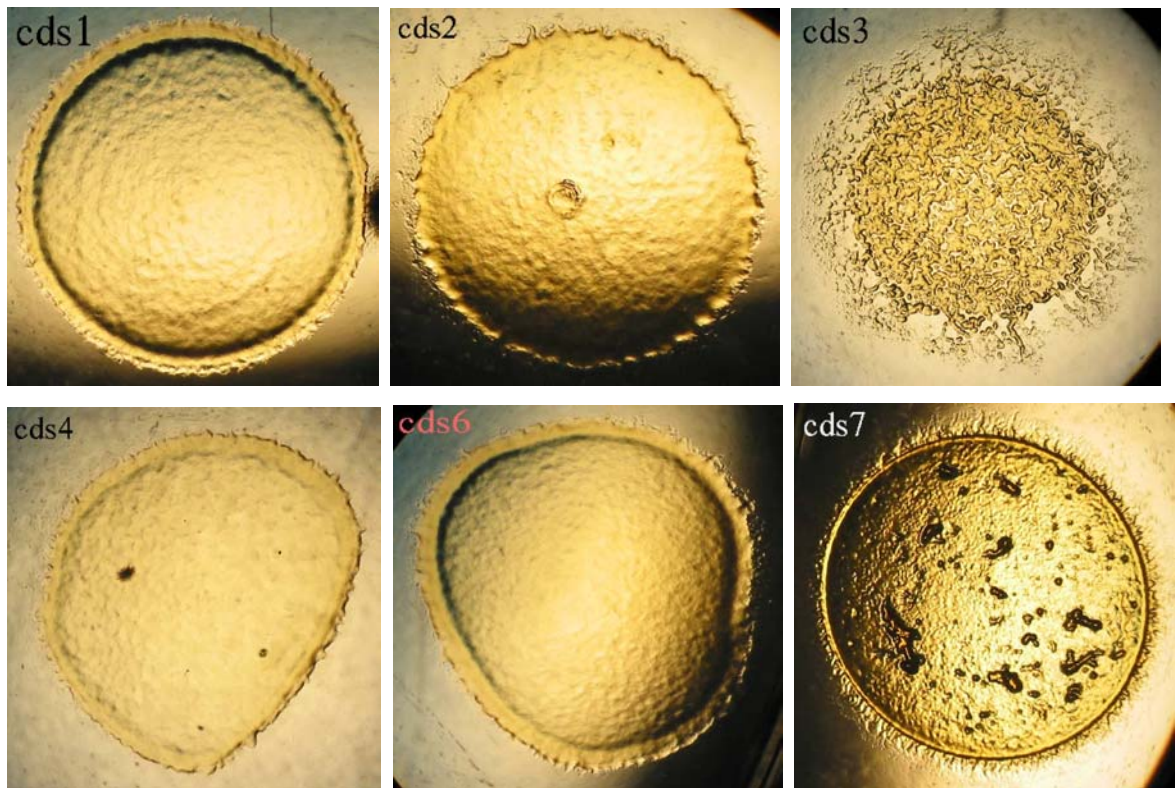


Figure 1.6 Developmental defects of mutants compared to the wildtype DK1622 and previously characterized strains SR53, SR171, SR200 and *esg*. Each spot contains 10 μ l of concentrated cells of the respective strains on developmental agar CF after 96 hours of incubation. (Continued on next page)

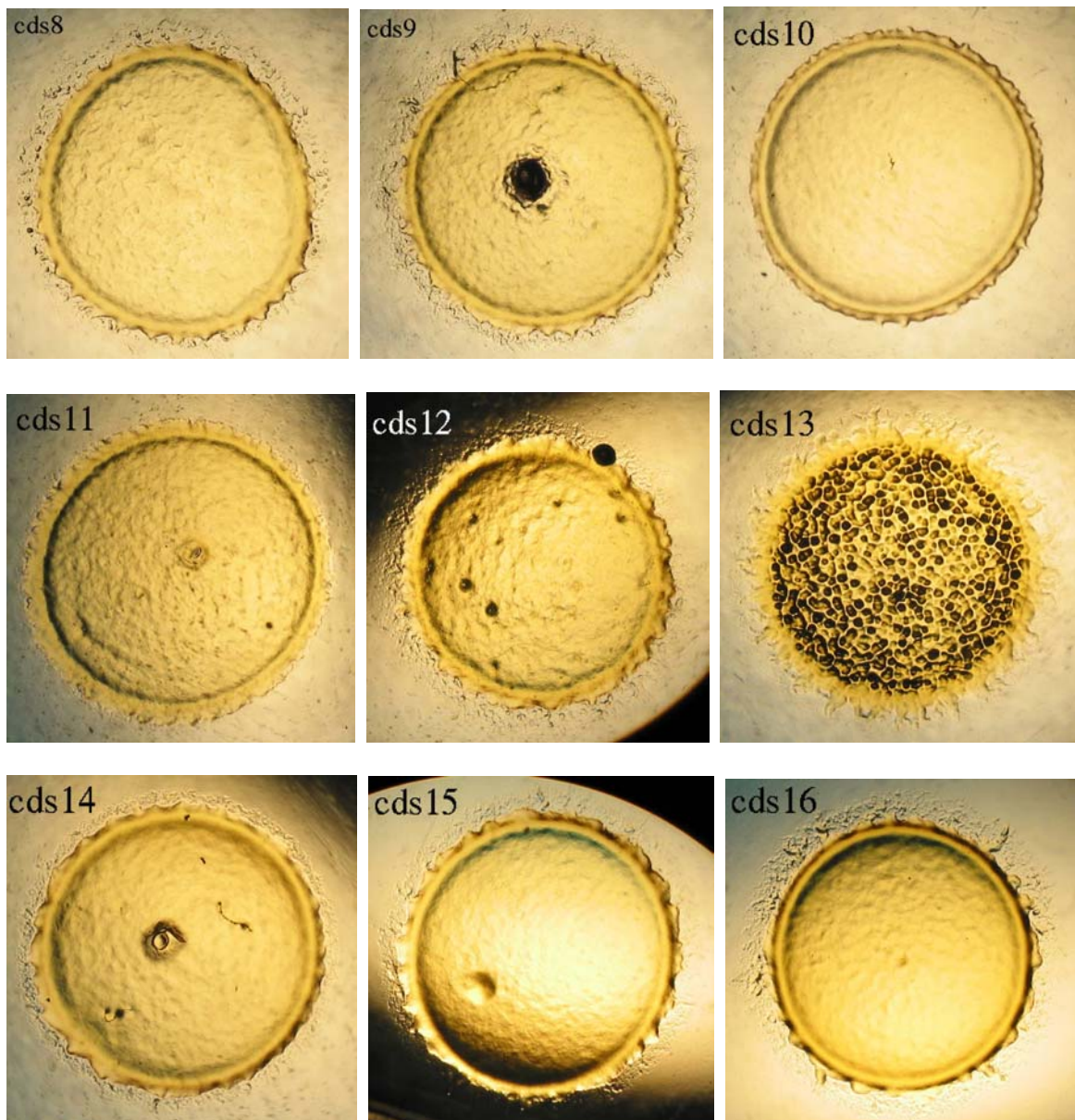


Figure 1.6 Developmental defects of mutants compared to the wildtype DK1622 and previously characterized strains SR53, SR171, SR200 and *esg*. Each spot contains 10 μ l of concentrated cells of the respective strains on developmental agar CF after 96 hours of incubation. (Continued on next page)

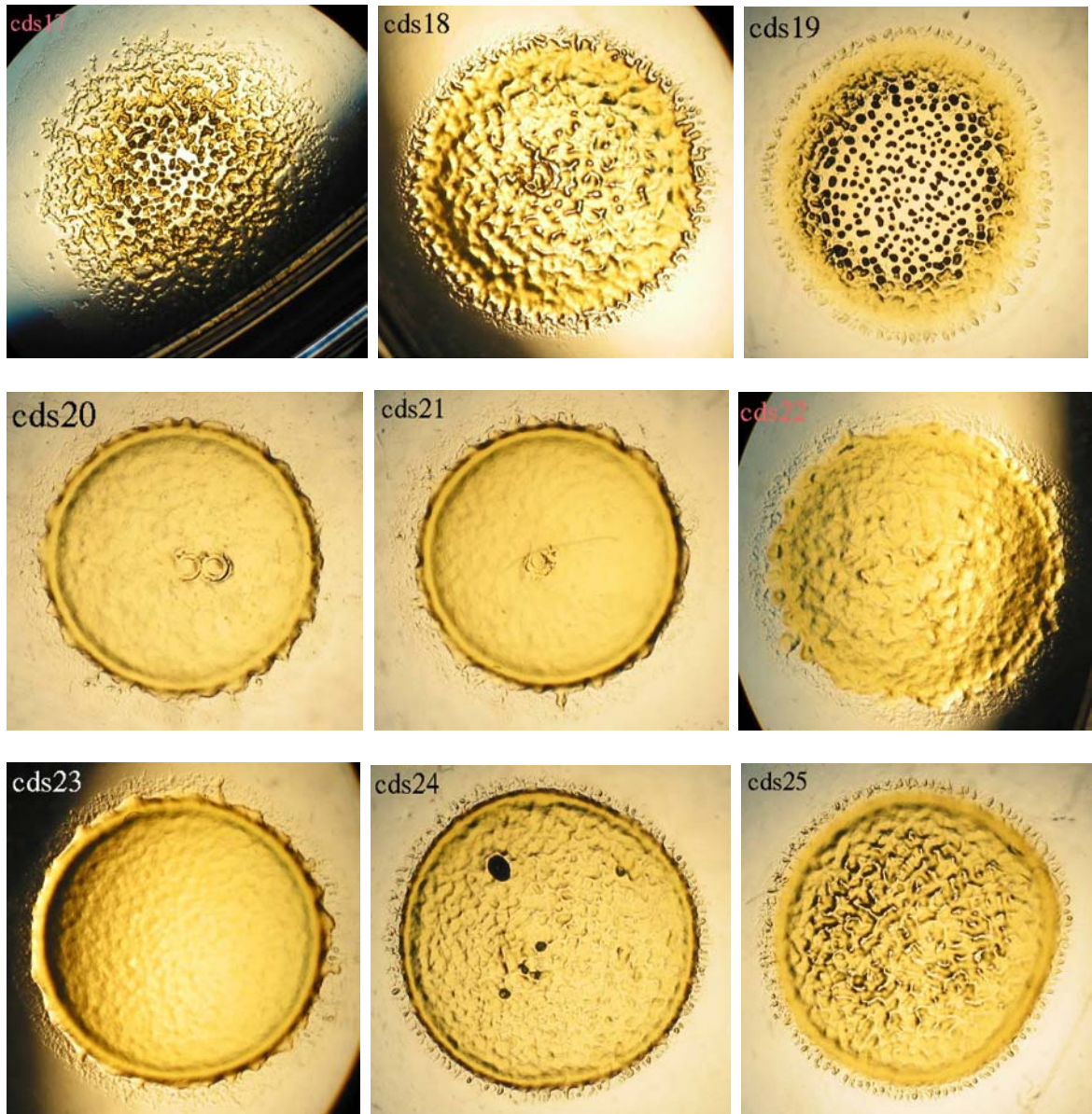


Figure 1.6 Developmental defects of mutants compared to the wildtype DK1622 and previously characterized strains SR53, SR171, SR200 and *esg*. Each spot contains 10 μ l of concentrated cells of the respective strains on developmental agar CF after 96 hours of incubation. (Continued on next page)

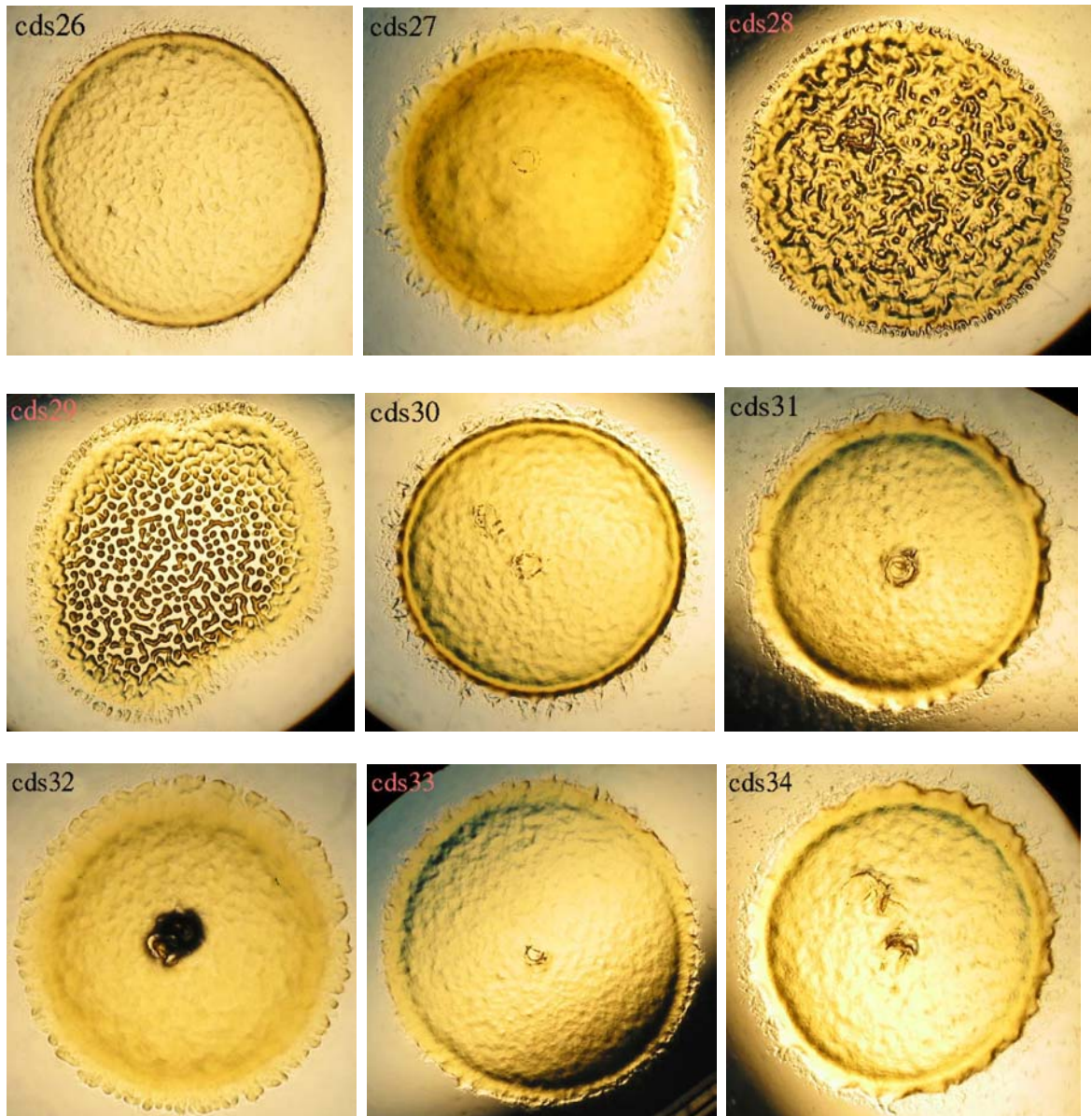


Figure 1.6 Developmental defects of mutants compared to the wildtype DK1622 and previously characterized strains SR53, SR171, SR200 and *esg*. Each spot contains 10 μ l of concentrated cells of the respective strains on developmental agar CF after 96 hours of incubation. (Continued on next page)

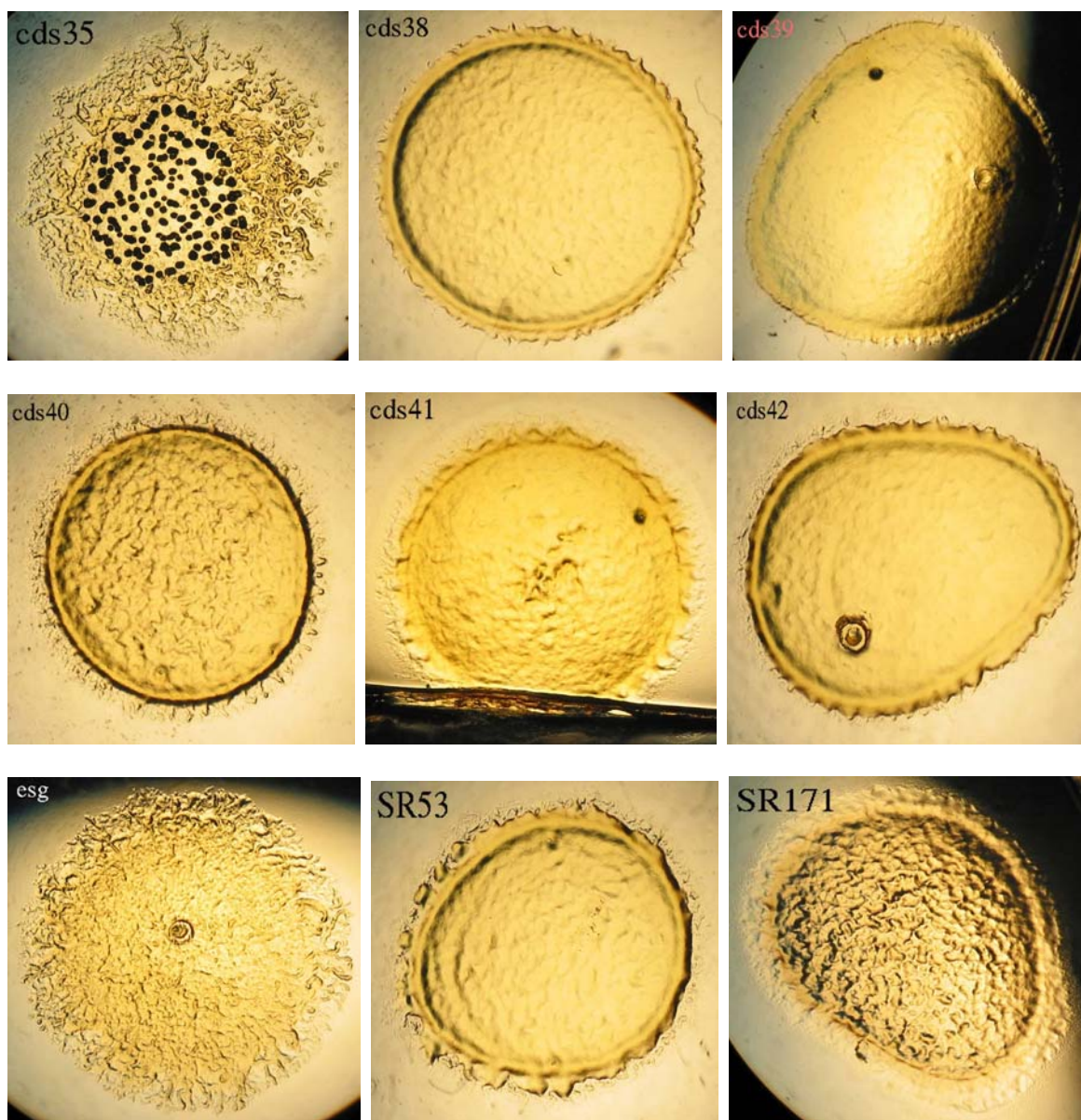


Figure 1.6 Developmental defects of mutants compared to the wildtype DK1622 and previously characterized strains SR53, SR171, SR200 and *esg*. Each spot contains 10 µl of concentrated cells of the respective strains on developmental agar CF after 96 hours of incubation. (Continued on next page)

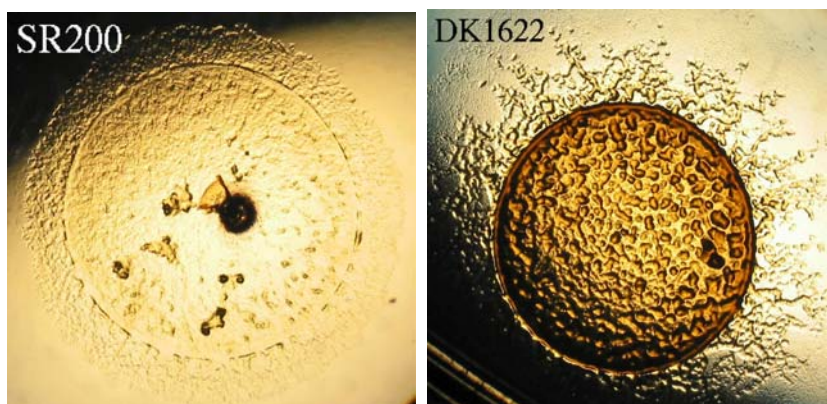


Figure 1.6 Developmental defects of mutants compared to the wildtype DK1622 and previously characterized strains SR53, SR171, SR200 and *esg*. Each spot contains 10 μ l of concentrated cells of the respective strains on developmental agar CF after 96 hours of incubation. This figure spans for six pages, see the previous pages for 1.6A-E.

The cohesion-proficient strains identified in the previous section tended to be less severely impaired in development. Strain *cds13*, *cds17* and *cds35* all formed fruiting bodies and sporulated (Figure 1.6). As an exception, the *cds27* seemed to be unable to aggregate and sporulate although its cohesion property did not show much deficiency. The developmental morphology of *cds27* is quite uniquely yellow, and thick, indicating that it has been growing on CF plate, instead of developing.

Calcofluor White-Binding Defects

Calcofluor white is believed to bind to nascent chitin and glucan on the cell surface and makes the cell fluorescent (Lussier *et al.*, 1997). The ability of binding Calcofluor White for each strain was assessed by spotting 10 µl of concentrated cells ($\sim 5 \times 10^9$ cells/ml) on low percentage of agar (0.3%), containing Calcofluor White (50 µg/ml). Fluorescence was monitored with a UV light source (366 nm). Wildtype cells were brightly fluorescent after 24 hours, and the mutants were either not fluorescent or very very slightly so. After observing these fluorescence patterns of cell spots in several experiments, it became clear that the pattern could be used to differentiate between the levels of fluorescence in the different polysaccharide production-deficient strains. This can be seen in the Figures 1.7 and 1.8 after 48 and 96 hours of incubation, respectively. The fluorescent ringed-structure was visible after 48 hours of development, and became very clear after 96 hours of development.

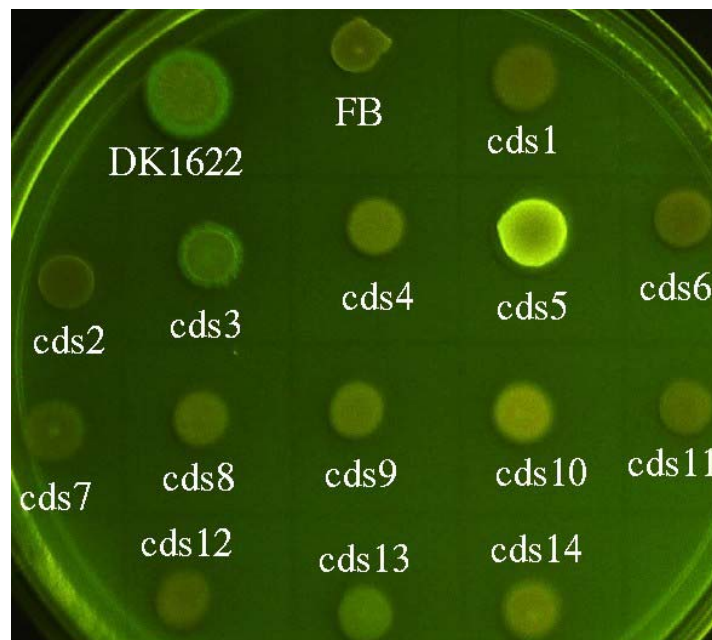


Figure 1.7 Calcofluor White-binding assay. Concentrated cells of each strain were spotted on the Calcofluor White-containing soft agar surface and incubated for 48 hours in 10 cm plates (nominal). Note: cds5 is contaminated in this picture (continued on next page)

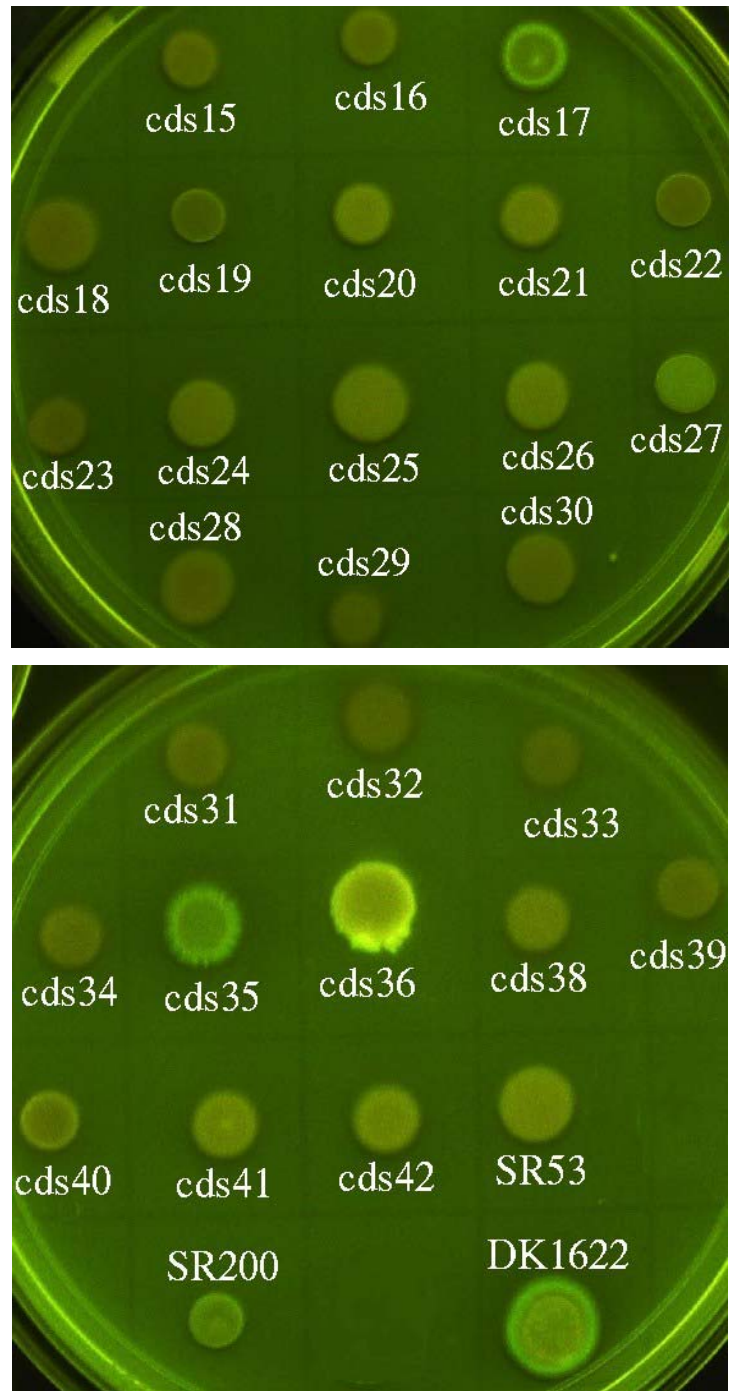


Figure 1.7 Calcofluor White-binding assay. Concentrated cells of each strain were spotted on the soft agar surface and incubated for 48 hours in 10 cm plates (nominal). Note that the spots of cds5 and cds36 were contaminated; therefore the fluorescence cannot be used to judge those strains. Fortunately, these two strains are not in our focus of study here.

The total fluorescence of each spot was significantly enhanced over time. For example, the wildtype DK1622 went through dramatic changes in its fluorescent pattern. After 48 hours, (Fig. 1.7) the entire spot was fluorescent. After 96 hours, however, (Fig 1.8) the center of the spot became darker, a very dark ring appeared surrounding that, and a very bright ring of fluorescence was present, with a radius larger than the original spot. The ring was probably composed of cells that had migrated or grown at the fringe in the ring. The wildtype behavior was the most dramatic of this system. However, most of the mutant strains gave a similar pattern of florescence over time, but with much less intensity of the spot. This is not very visible in the Fig. 1.8 due to the loss of fidelity of the printing system. The original photographs show the banding patterns much more clearly.

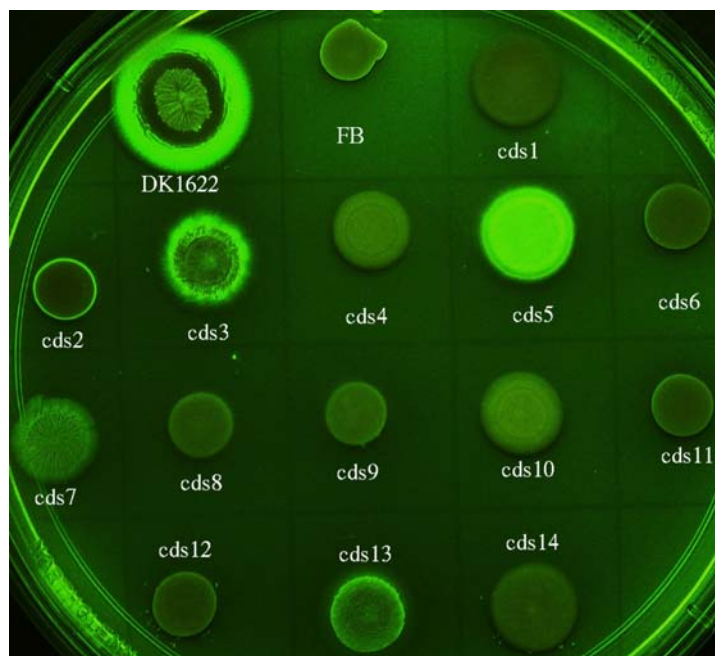


Figure 1.8 Enhanced Calcofluor White-binding after 96-hour incubation at 30°C in 10 cm plates (nominal). With enhanced fluorescence, the level of the Calcofluor White-binding deficiency for each mutant is clearly distinguishable. See text for details. Note: cds5 on this plate is contaminated. (Continued on next page)

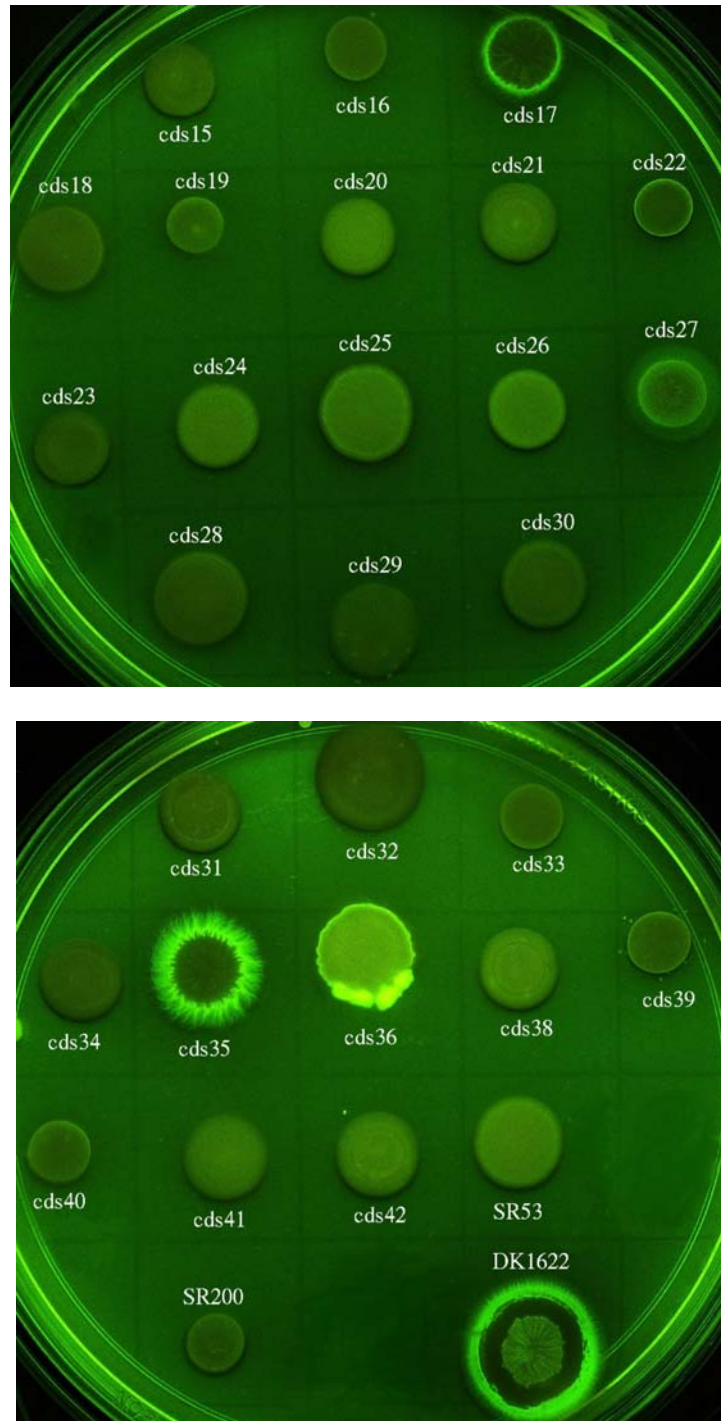


Figure 1.8 Enhanced Calcofluor White-binding after 96-hour incubation at 30°C in 10 cm plates (nominal). With enhanced fluorescence, the level of the Calcofluor White-binding deficiency for each mutant is clearly distinguishable. See text for details. Note: the cds36 is contaminated.

With the enhanced fluorescence, it is clear that the fluorescence level between the mutants could be distinguished (Fig. 1.8). Upon careful examination of the changes of the fluorescence in each spot, the fluorescence enhancement is not a simple increase of brightness, but a complex sum of some parts becoming brighter, others becoming darker under ultraviolet light. The spatial distribution changes generally follow a simple rule: the inner part of the spot becomes darker while the outer part becomes brighter. This observation seems to agree with earlier reports (Steer, 1977; Lussier *et al.*, 1997). For example, the wildtype strain DK1622 is fluorescent through out the whole spot after 48 hours of spotting. Over time, the interior of the spot gradually loses fluorescence, even become darker than the medium, whereas the growing fringe's fluorescence becomes increasingly brighter. This is generally true for most mutant strains too.

Some mutants display even more complex fluorescence patterns. For example, *cds3* has a bright fringe and fluorescent rings in the center of the spot, even some radial patterns are visible. The similar fluorescence pattern was observed in the strains *cds13*, *cds17*, and *cds35*. The strain *cds4* was slightly fluorescent throughout the whole spot after 96 hours of incubation, with a thin ring of darkness marking the size of the original spot (Fig. 1.8, p.41). The fluorescence pattern of *cds10*, *cds20*, *cds28*, *cds41* and *cds42* was very similar to that of the *cds4*, except that they were slightly brighter than *cds4* and had a larger radius fringe. Interestingly the strain *cds31* showed the opposite pattern. The whole spot was darker, while the ring marking the original spot size was brighter in fluorescence (Fig. 1.8, p.42). Strain *cds34* displayed a similar pattern to that of *cds31*. The *cds2* had a thin, solid, smooth, very bright spot edge. The center of the spot was completely dark (Fig. 1.8, p.41). This fluorescence pattern was shared to a variable degree by many smooth-looking strains,

including FB, cds6, cds8, cds9, cds11, cds12, cds16, cds19, cds22, cds33, cds39, cds40 and SR200. Another pattern was uniformly dimly fluorescent throughout the spot, including cds1, cds14, cds15, cds18, cds21, cds23, cds24, cds25, cds26, cds28, cds29, cds30, cds32, and SR53. Strain cds1 and cds32 were the least fluorescent ones in this set. These fluorescence patterns are relatively difficult to observe if the fluorescence pictures are not taken in the complete darkness.

Table 1.3 Calcofluor White-binding fluorescence patterns in the test spot after 96 hours of incubation for strains on 0.3% agar containing CYE and Calcofluor White.

Pattern	Description	Strains
1	fluorescent throughout the whole spot after 40 hours of spotting. Over time the interior of the spot gradually loses fluorescence, even become darker than the medium, whereas the growing fringe's fluorescence becomes increasingly brighter	DK1622
2	A wide and bright fringe and fluorescent rings in the center, even some visible radial patterns.	cds3, cds13, cds17, cds35
3	The fringe is darker than the inner part of the spot, but the center is darkest.	cds27
4	Slightly fluorescent throughout the whole spot after 96 hours of incubation, with a thin ring of darkness marking the size of the original spot	cds4, cds10, cds20, cds28, cds41, cds42
5	The whole spot is darker, the ring marking the original spot size is much brighter in fluorescence	cds31, cds34
6	A thin, solid, smooth, fluorescent border, no fringe. The center of the spot is completely dark.	cds2, cds6, cds8, cds9, cds11, cds12, cds16, cds19, cds22, cds33, cds39, cds40, SR200, FB
7	Uniformly, dimly fluorescent throughout the spot	cds1, cds14, cds15, cds18, cds21, cds23, cds24, cds25, cds26, cds28, cds29, cds30, cds32, SR53

The level of fluorescence in the patterns 3 through 7 was dramatically lower than the wildtype. Although the patterns are clear on computer screen, are difficult to show clearly in printed form.

The results of cohesion tests and Calcofluor White-binding experiments on the mutant strains seemed to be consistent with each other. For example, cohesion proficient strains (eg. DK1622, *cds3*, *cds13*, *cds17*, *cds27*, and *cds35*; refer to Fig 1.4) also showed higher Calcofluor White-binding than the cohesion deficient, with the *cds27* showing the least fluorescence intensity. At this point, the meanings of the fluorescence patterns are not known. According to the pattern changes in the wildtype, the bright fluorescence seemed to emit from the freshly growing cells at the fringe. When the cells grew older (like the ones inside the spot), they appeared to gradually lose fluorescence, at the same time as the growing fringe became extraordinarily bright.

Searching The *M. xanthus* Genome Database To Assemble “Retrieved Contigs”

To find out where these insertions are in the genome, a genomic sequence map has to be established. Fortunately, the *M. xanthus* genome is being sequenced, first at the firm Cereon, and subsequently at The Institute for Genomic Research (TIGR), funded by National Science Foundation. In the early days of this project, the retrievable sequence length was strictly limited by the servers. To put the insertions in a broader context of the genome, it is necessary to re-assemble the retrieved fragments into “retrieved contigs”, Rcontig. The most basic function for the retrieved contigs is to rejoin the fragments of database contigs retrieved. Then the Rcontigs are stored in our own database. However, the Rcontigs are only intermediate contigs, when the database contigs gradually integrated into longer ones, retrieved contigs were merged into newer, longer database contigs. Obviously once the whole genome sequence is finished all contigs become one, and all insertions will be on a single “contig” – the complete genome.

Insertion Point Sequencing and Mapping

At the time when this project started we knew we could use a *M. xanthus* genomic database. Therefore, we determined that the insertion point sequence data does not have to achieve high precision. Each insertion was sequenced only once from each end of the transposon. We call this “single pass sequencing”. Single pass sequencing sometimes yielded very unpredictable sequence qualities. Compared to the *M. xanthus* database at NCBI (National Center of Biotechnology Information), our insertion point sequences have a range of differences: from no error in more than 600bp (SR53_left) to only a small part of our own sequence having a credible match to the database counterparts (Table 1.4, see the Figures 2.14 and 2.13) (error rate >5%). This creates problems for insertion site mapping. Normally when we talk about comparing two sequences, we tend to think about two finalized high quality sequences. Therefore, any differences between the two will signify their actual differences between two pieces of DNA. However even today in the year 2004, it is still difficult, time consuming, and expensive to always have high quality sequences before starting sequence analysis work, especially when there are many sequences to be processed. Insertion point mapping is particularly so. When using high quality sequences, one can use any word processing program, such as Notepad, and search for a tiny fragment of one sequence against the other to find out where the two sequences start to differ, and then mark the insertion point there.

Table 1.4 Single Pass Sequencing Error Rates (A sample). Just some examples.

Flanking Sequence	Leading*	Matching*	Tailing*
SR53 left	6/38	0/630	>100/370
cds21 mar1	4/29	53/420	0/0
cds21 mar2	4/20	134/460	0/0
cds1 mar1	6/30	1/374	20/22
cds1 mar2	5/28	4/374	20/22
cds19 mar1	4/20	9/602	6/21
cds19 mar2	0/0	0/555	>100/341

* The error rate in a region on the sequence that matched the database sequence is listed under “matching”. The error rate in the region before the “matching region” is listed under “leading”. The error rate in the region after the “matching region” is listed under “trailing”.

When the available sequences are known of low precision, as long as one of the pair is of low quality, the approach above would not work. It is for this reason that a special computer program was started. The computer program is designed to be able to take into account the sporadic errors that may occur in a DNA sequence, and generate a visual image to report the result. Figure 1.9 shows an example. See Chapter 2 for details. This program is especially useful when the sequences one wants to search for and match the query sequences to is not in a ready made BLAST-searchable database. That was the initial situation I was in when I started the sequence analysis of this set of insertions.

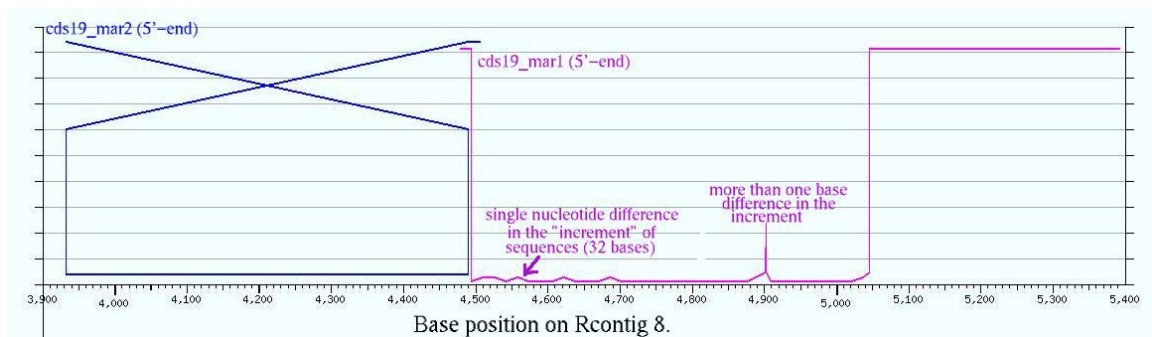


Figure 1.9. Insertion cds19 maps to the Rcontig8, at base position 4490. The primer mar1 is oriented in the same direction as the contig. The primer mar2 is oriented in a reverse direction, therefore the line representing the query sequence derived from mar2 has a cross in the figure.

How to read the mapping result: Take the cds19_mar1 as an example. The little horizontal purplish line (about 20 bp) on the top represents a very small piece of 5'-sequence does not match the Rcontig8. Then the vertical line simply means start from that base position (~4490 bp) on the Rcontig8 the two sequences have significant homology. Since the homology is not 100%, the bottom line is not straight for ~200 bp, then the two sequences matched 100% for about 180 bp, then comes another region of disagreement at around 4900 bp position on the Rcontig8. It is followed by a 100 bp stretch with 100% agreement. Towards the end for about 350 bp, the two sequences do not match any more, therefore the query sequence is lifted up high. In short, the aligned part of each query sequence is the segment that is close to and parallels the X axis. Different sizes of bends in the aligned segment represent various degrees of errors in those parts of the query sequence. (see Chapter 2)

Assembling The Genomic Sequence Map

In searching for a way to put the insertional mutations into a broad context, we realized very

recently that not only the insertions could be mapped to the database contigs, but also the contigs in the *M. xanthus* genome database were long enough to be assembled into a genome sequence map. Therefore it was possible to establish a platform for discussing the mutations in the context of the whole genome. The genomic sequencing of *M. xanthus* is more than 95% completed. However the genomic sequence is in more than 40 pieces (contigs). Their relative location and orientation on the genome is not known yet. A previously published physical map of *M. xanthus* (He et al., 1994), and sequences from published genes were used to construct a genomic sequence map (Figure 1.10).

The procedure is very similar to a plasmid restriction map constructed from a set of restriction enzyme digestions. A simple computer program was designed to find the restriction sites and positions for SpeI (ACTAGT) and AseI (ATTAAT) in each of the contigs. SpeI and AseI were chosen because they were the two restriction enzymes used in previous physical mapping studies (Kuspa et al., 1989, He et al. 1994). The assembly of this genomic sequence map was possible because the available contigs are long enough to intersect with previously mapped restriction sites. The computer program searches the contigs and returns the restriction site's base positions. From the base positions the intervals are calculated and compared with the sizes of the physical map.

A fragment from the computer search that matches a fragment on the physical map is considered the same fragment on the physical map of the genome. The neighboring fragments relations, or the sequence of the restriction sites intervals determines orientation of the contigs on the genome. Repeating this process for all contigs, and the whole genome

sequence map is constructed (Fig. 1.10).

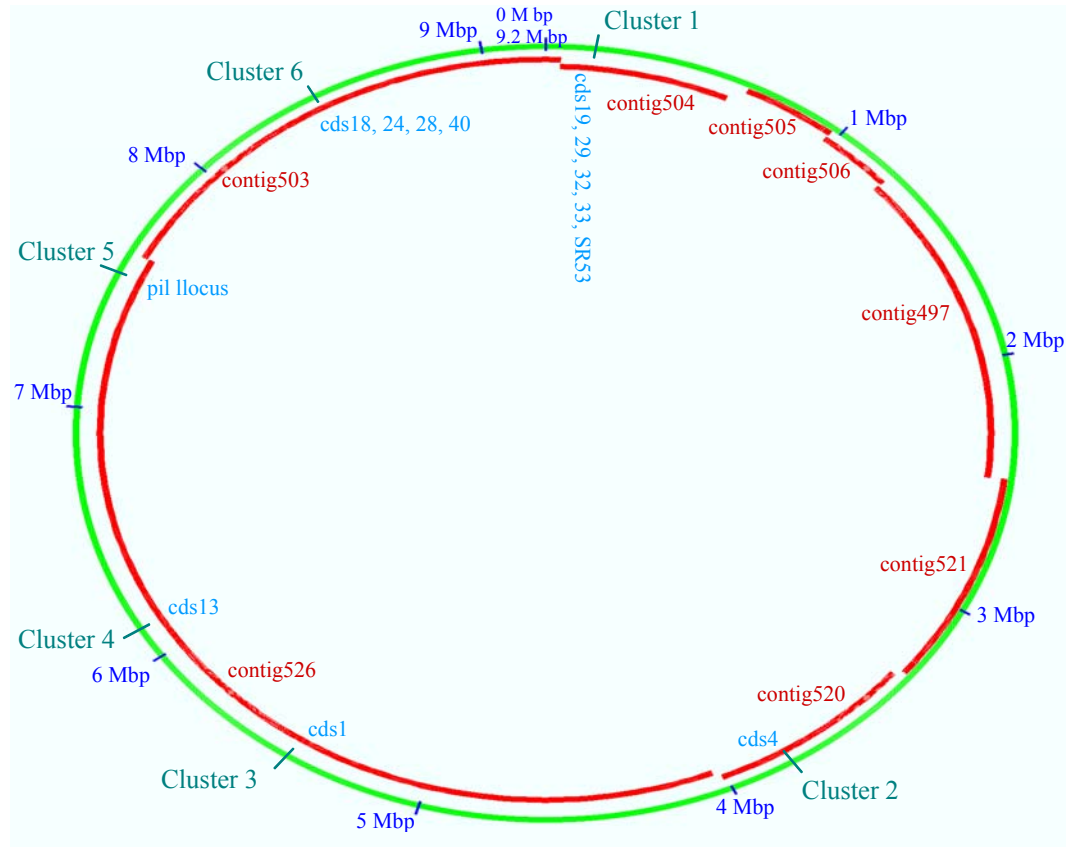


Figure 1.10 Genomic sequence map assembled with an unconventional technique, which is similar to construction of a plasmid map structure from restriction enzyme digestion – gel electrophoresis results. This technique circumvents the need to know the complete genomic sequence in order to construct a genomic sequence map. Short blue radial lines mark the sequence position according to the physical map (He et al. 1994). Red arcs are the contigs. Thin blue radial lines represent the transposon insertion clusters.

The eight long contigs (red arcs in Figure 1.10) that carry the restriction sites were mapped. Although there are more than 40 contigs in the NCBI unfinished *M. xanthus* genome, these eight contigs however represent more than 95% of the total available sequence length in the *M. xanthus* genomic sequence database. This means that no possibility exists for any

conflict when the whole genome is completely sequenced, unless some of the eight contigs used are to be found seriously defective. By the way, many of those small contigs are duplicates of parts of the 8 long contigs. Therefore, I believe, many of the small contigs will eventually be eliminated upon enough evidence to prove their redundancy.

Through this genomic sequence map reconstruction, a few extra *SpeI* restriction sites were found. These extra site are ones not mapped on the published physical map of *M. xanthus*. However these additional sites generate very small fragments, usually about one kilobase or less in length. Therefore, their discovery does not change the overall structure of the physical map. We did find inaccuracy on the published physical map. For instance, the gene *csgA* was mapped to 2.3 Mbp position, whereas its sequence homology shows it should be mapped to 1.7 – 1.8 Mbp region. This can be easily explained considering the inability for placing a restriction fragment in a right orientation when there is no restriction sites exist within that fragment. In general, the accumulated *Myxococcus xanthus* genomic sequence agrees with the published physical map very well.

Genetic Map Of The Insertion Sites

Transposon mutants were generated by electroporating the *magellan-4* carrying plasmid into the wild-type *M. xanthus* DK1622 and selecting for kanamycin resistant and Calcofluor White-binding deficient colonies as described in the material and methods. A total of 43 mutants were selected from thousands for this analysis. Due to time and resource restrictions, only 26 of them were sequenced, 21 unique insertions were mapped. Two insertions were collected and sequenced more than once.

The insertion point sequence analysis is continuously evaluated as new data becomes available in the GenBank database. At this point, large contigs have been assembled surrounding all insertion sites. Each of these sites encompasses tens of genes probably transcribed in many different operons. Although we annotated almost all genes in each of these clusters, many of them are not directly related to the insertion mutations (not shown). Therefore, only described are those ORFs that are transcribed in the same transcription unit as one of those insertions in this chapter. Because there is no easy way to know the actual transcription unit, these possible transcription units are called “apparent operons” here for convenience. Sequence analysis results in the following were based on BLASTP search results against the NCBI non-redundant databases with all default settings unless specifically noted otherwise. In addition, domains refer to the domains found in BLASTP search, and score and E values refer to, respectively, the score value and E value found in the BLASTP search results.

Table 1.5 The relationship between the insertions, the Rcontigs and the NCBI database contigs. Mutants cds23 and cds27 were the same insertion. Mutants cds14, cds20, and cds37 were found to be the same insertion.

Mutant (Insertion)	Rcontig*	Contig in NCBI database
cds1	1	526
cds4	2	520
cds8	4. 5. 6	526
cds9	4. 5. 6	526
cds11	4. 5. 6	526
cds13	36	526
cds14 (cds20. cds37)	4. 5. 6	526
cds16	4. 5. 6	526
cds18	27	503
cds19	8	504
Cds21	15	526
cds22	16	526
cds23 (cds27)	15	526
cds24	18	503
cds28	27	503
cds29	21	504
cds32	22	504
cds33	7	504
cds40	18	503
cds42	16	526
SR53	29	504

- These Rcontigs are the intermediary contigs for the mapping process only. As the genomic sequencing progressed, and the retrieved sequences extended, Rcontigs gradually merged with the NCBI database contigs. Eventually when the genomic sequencing finishes, every cds insertion will be on the single contig, representing the whole genome. At the time of this writing, the genome sequencing is still unfinished. The insertions' location on the genome is presented in the Figure 1.10.

Cluster 1 (Figure 1.11)

Cluster 1 is at 0.13 Mbp position on the physical map of *M. xanthus* (Figure 1.10). Insertion SR53 was created earlier by Ramaswamy (Ramaswamy *et al.*, 1997). It is described as one of the classical cds (Calcofluor White-binding deficient and S motile) strains (Ramaswamy *et al.*, 1997). Now the SR53 locus has been cloned, sequenced and mutagenized to further characterize the cds phenotype. Within the apparent operon of SR53 there are three ORFs orfC1-2-5, orfC1-2-6, and orfC1-2-7 (Fig. 1.11). The insertion SR53 is in the orfC1-2-5. orfC1-2-5 has a strong homology to a glycosyltransferase from *Rhodobacter sphaeroides* (score 110, E $7e^{-23}$). In addition, it matches conserved domains: (1) COG0438, RfaG, Glycosyltransferase (Score 117, E $3e^{-27}$); (2) pfam00534, Glycos_transf_1, Glycosyl transferases group 1 (Score 101, E $2e^{-22}$); Mutations in this domain may lead to disease in humans (Paroxysmal Nocturnal haemoglobinuria). Some members of this family transfer activated sugars to a variety of substrates, including glycogen, Fructose-6-phosphate and lipopolysaccharides. Others transfer UDP, ADP, GDP or CMP linked sugars. The eukaryotic glycogen synthases may be distant members of this family.

The phenotypic characteristics of this insertion mutation SR53 agree with the predicted function of the gene as a glycosyltransferase. Strain SR53 produced much less polysaccharide, less as indicated by Calcofluor White-binding. However, this strain retains a substantial amount of S-motility (Ramaswamy *et al.*, 1997). It is now clear that the S-motility in the cds mutants varies to the full range (Figs. 1.4, and 1.5; Table 1.2). Predicted ORFs marked with red arrows to the left of SR53 were targeted for further study. Therefore they are discussed below in the section on “Internal Replacement Mutagenesis Analysis

of the Two Selected Operons”.

The orfC1-2-6 has good homology to the two-component hybrid sensor and regulator VicK (score 151, E $4e^{-37}$) known in *Gloeobacter violaceus*. The orfC1-2-6 also has numerous matches in the conserved domain databases. For example, it is homologous to (1) cd00156, REC, Signal receiver domain with a score 43.3 and E value $8e^{-06}$; (2) pfam00072, response_reg, a response regulator receiver domain with a score 41.0 and E value $4e^{-05}$. (3) smart00448, REC, cheY-homologous receiver domain with a score 40.6 and E value $5e^{-05}$.

The last ORF orfC1-2-7 is highly homologous to many domains, the best homology is with the *mgtE* from *Bdellovibrio bacteriovorus* (score 386, E e^{-106}). This is a CBS-domain-containing membrane protein domain, part of the signal transduction mechanisms found in *Pseudomonas aeruginosa* and *Agrobacterium tumefaciens* among many others. The best whole protein homology is found in the Mg^{++} transporter from *Xanthomonas axonopodis* pv. citri str. 306 (score 281, E $2e^{-73}$).

The results from this apparent operon seem consistent with the phenotype observed for SR53. Loss of glycosyltransferase leads to defects in polysaccharide production. A recurring feature of glycosyltransferase operon arrangement is its being interspersed with two-component signal transduction genes. Since the SR53 insertion is in the upstream ORF, the phenotype could be due partially to polar effects. But the observed significant loss of Calcofluor White-binding is consistent with the loss of orfC1-2-5 glycosyltransferase.

Insertion *cds19* is in *orfC1-2-9*, only one ORF away from the *SR53* operon. This seems to be a monocistronic operon. The *orfC1-2-9* is very similar to *orfC1-2-5* that *SR53* is in. It also has a strong homology to RfaG (score 125, $E\ 4e^{-29}$). Its most probable function is as a glycosyltransferase. The growth and developmental defects are much more mild than those of *SR53*. This is probably because the *orfC1-2-9* is a monocistronic operon. Most of the developmental defects in *SR53* could be due to the polar affects on the downstream genes. Loss of *orfC1-2-6* (a putative histidine kinase, a sensor in the two component signal transduction system) and *orfC1-2-7* (a potential Mg/Co/Ni transporter) is conceivably serious to the development process.

The two apparent operons flanking the *cds19* operon are also small. The one upstream from *cds19* operon is another monocistronic operon transcribed in divergent direction, encoding a protein with good homology (Score 59.7, $E\ 4e^{-07}$) with a hypothetical protein from *Chloroflexus aurantiacus*. The one downstream from the *cds19* operon is a dicistronic operon, transcribed in the convergent directions. One of them encodes an NtrC-like activator known in *M. xanthus* (Caberoy *et al.*, 2003). The other encodes an enzyme, MoeA, a molybdopterin biosynthesis enzyme from *Geobacter metallireducens* (score 218, $E\ 5e^{-55}$).

About 22 kbp downstream from the *cds19* are the insertions *cds29*, *cds32*, and *cds33*, which are in the same apparent operon, consisting of ten genes, transcribed in the reverse direction. Two of them are glycosyltransferases, *wcaA* (*orfC1-5-23*) and *wcaJ* (*orfC1-5-28*). Another three of them are membrane proteins involved in export of polysaccharide (*orfC1-5-27* and

orfC1-5-24) or its derivatives (orfC1-4-34). At the leading end of the apparent transcript are two genes related to molybdopterin biosynthesis, *moeA* and *fdhD-mobB*. Downstream from the *cds29* insertion there is a small two-component signal transduction protein orfC1-4-31. The ORF between the *cds29* and the two-component signal transduction protein is an ORF, orfC1-5-26, that has no database match. Downstream from the 2-component signal transduction ORF are two ORFs involved in polysaccharide biogenesis orfC-1-5-24 (outer membrane protein) and orfC1-5-23 (*wcaA*). These two ORFs together with the orfC1-5-26 were further investigated using internal fragment replacement mutagenesis (See below). At the end of the transcript is an ORF orfC-1-4-29 matching a hypothetical protein from *Mycobacterium tuberculosis* (Score 108, E $7e^{-23}$). However, this ORF has a low G+C bias (75.2%) at the third base in the codons.

The insertions *cds32* and *cds33* both are in the third gene (orfC1-5-28) on the transcript, which has a very strong homology (score 207 E4e-52) to *wcaJ*, a glycosyltransferase from *Geobacter sulfurreducens*. The insertion *cds29* is in the fifth ORF on this transcript, which encodes a outermembrane or periplasmic protein involved in polysaccharide export (score 79, E $4e^{-13}$).

To summarize, among the potential ORFs on this apparent transcript are a set of three genes that are putatively involved in polysaccharide production: orfC1-5-27, orfC1-5-24, orfC1-5-23, and two others belonging to regulatory systems, orfC1-4-31 and orfC1-4-29. The three glycosyltransferase ORFs are highly homologous to known database sequences found in other organisms. The insertion *cds29* is in the orfC1-5-27. The orfC1-5-27 hits two conserved domains: pfam02563 poly_export, a family of periplasmic proteins involved in

polysaccharide biosynthesis and/or export; and COG1596 Wza, a periplasmic protein involved in polysaccharide export. The orfC1-5-27 is highly homologous to many known genes. A characteristic feature of these homologs is a periplasmic protein known to be involved in polysaccharide export in *Geobacter metallireducens* with a score of 78.2 and E value of $5e^{-13}$.

The upstream ORF in this long chain of ORFs is the orfC1-4-36. It contains a conserved domain MoeA (Score 323, E $9e^{-89}$, CD length 404), which is involved in molybdopterin biosynthesis. The domain contains two subdomains MoeA_N and MoCF_biosynth. MoeA is a molybdopterin biosynthesis enzyme. Its exact function is not clear. Note that this is the second *moeA* gene in this cluster. The other *moeA* is second downstream from cds19 insertion.

The orfC1-5-29 has a coding capacity of 588 residues, overall third base G+C ratio of 73%. The first 122 codons are almost completely composed of low complexity sequence fragments, have a third base G+C bias (49%) far below the overall G+C ratio for the coding capacity, are overlapped by an upstream ORF orfC1-4-36 *moeA* on the transcript. If these 122 codons were excluded, the remaining coding sequence has a third base G+C ratio at 91%. This will create a putative peptide with a four-base overlap at the N-terminus with orfC1-4-36 upstream. This overlap is considered highly likely because otherwise the N-terminal two thirds of the ORF would not have a methionine. Therefore, the peptide coded in orfC1-5-29 is estimated to be 466-residues long.

Another feature of this peptide orfC1-5-29 is that it contains two domains belonging to two completely different kinds of proteins. The N-terminus domain is MobB. The MobB domain COG1763 is 161-residues long, has a homology at Score 110, and $E = 7e^{-25}$. It is a molybdopterin-guanine dinucleotide biosynthesis protein, involved in coenzyme metabolism. The C-terminus domain of orfC1-5-29 is homologous to FdhD. The FdhD domain COG1526 is 266-residues long, has a homology at Score 210, $E = 8e^{-55}$. This domain is derived from some uncharacterized proteins required for formate dehydrogenase activity involved in energy production and conversion. This carboxyl domain is also similar to FdhD-NarQ. It is said that the NarQ part is a nitrate assimilation domain, and both FdhD and NarQ domains are required for formate dehydrogenase activity.

This compact and complex combination of domains in orfC1-5-29 is very interesting in several aspects. First, the NarQ and FdhD combination occurs in many organisms, such as *Mycobacterium tuberculosis*, *Corynebacterium glutamicum*, *Xanthomonas axonopodis*, etc. It might be valuable for evolutionary studies between various organisms. Second, the unique combination of MobB and FdhD-NarQ found in *Myxococcus xanthus* might indicate an enzyme's evolution in progress of a new protein and probably a new function. Third, the physiological relationships between these enzymes may be as close as they can be in a single peptide. Fourth, the regulatory mechanisms of these proteins in different organisms could be linked through that of orfC1-5-29. Consequently, this knowledge would lead to a better understanding the regulation of polysaccharide production. Finally, it would be interesting to know what properties of these domains make them to be more “malleable” than most domains. Or, is it the genome size of the host organism (*Myxococcus xanthus*'s

genome size is ~9.2 Mbp) that makes it more likely for new functions to develop?

The ORF orfC1-5-28 contains the *magellan-4* insertions cds32 and cds33. The carboxyl half of orfC1-5-28 carries two nested domains: COG2148 (Score 213 E $7e^{-56}$, over a 226-residues conserved domain) and pfam02397 (Score 185 E $1e^{-47}$, over a 197-residues conserved domain). WcaJ is a sugar transferase, located in the outer membrane, involved in lipopolysaccharide synthesis and cell envelope biogenesis. COG2148 is conserved in WcaJ proteins and proteins such as CpsD of *Bifidobacterium longum* (score 196, E $9e^{-49}$), Cps2E of *Pirellula sp.* (score 182, E $9e^{-45}$), and CpsE of *Streptococcus agalactiae* (score 182, E $1e^{-44}$). The domain pfam02397, nested inside the COG2148, is conserved in a number of different bacterial sugar transferase proteins, which are involved in diverse biosynthesis pathways. The closest homology (Score 207 E $4e^{-52}$) is found in a database protein from *Geobacter sulfurreducens*. Since the protein was found in a genome-sequencing project, its experimental nature is not known at this time. The orfC1-5-28 has a coding capacity of 510-residues, and the third base G+C ratio is 93% for the entire coding capacity. Its first methionine is at the residue 20. This makes the size of the protein to be 491 amino acids long, comparable to some members of the WcaJ family. However, the homology is only for the C-terminal one third to one half among members of this family. The N-terminal part is highly variable. Many members don't even have much more than the homologous region. About 110 codons at the N-terminal part of orfC1-5-28 are almost completely composed of low complexity sequences, and do not have any homology to any sequence in the GenBank.

The observed phenotypes of the insertions cds32 and cds 33 match the predicted gene

functions. Both insertions *cds32* and *cds33* cause smooth colony morphology, and deficiency in cohesion, Calcofluor White-binding and S-motility (Figs. 1.5, 1.6, 1.7 and 1.8). Although the insertions *cds32* and *cds33* are in the same ORF *orfC1-5-28*, the phenotypes of *cds32* and *cds33* are not completely the same. The *cds32* has a much better S-motility (65% of the wildtype) than the *cds33* (14% of the wildtype). However, it is the *cds33* that showed some residual Calcofluor White-binding capacity. There are other subtle differences between the two in growth, cohesion, and development as well. According to their insertion positions in the ORF, *cds32* breaks (7-8 amino acids) more off the glycosyltransferase *WcaJ* than *cds33* does. This seems to be the only genetic difference causing the motility change.

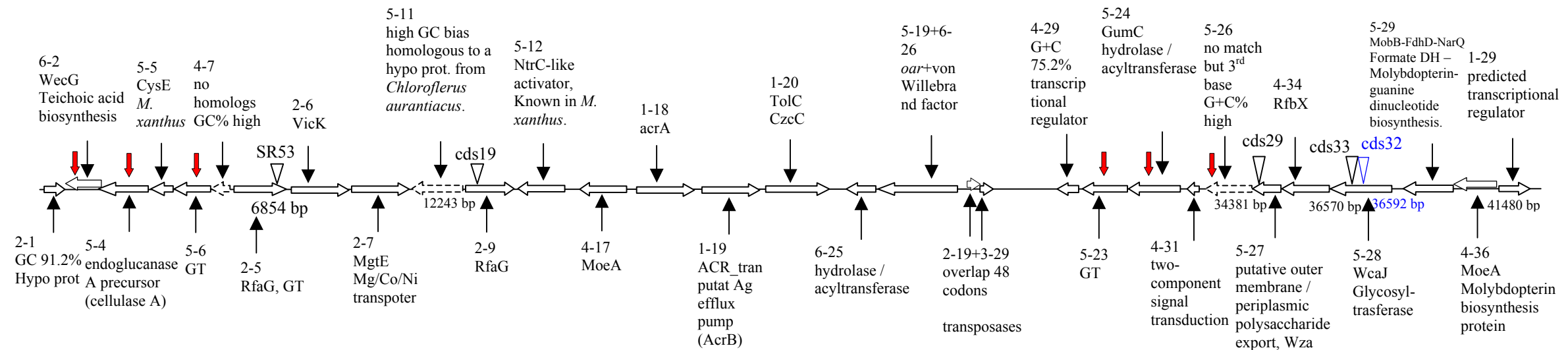


Figure 1.11 Insertion map of Cluster 1. dash-line-arrowed ORFs don't have database matches, but have high G+C third base codon bias. Red arrows indicate the sites for the PCR targeted mutagenesis analysis, which are discussed under the "Internal fragment replacement mutagenesis" at the end of Results section.

Abreviation: 2ST – two component signal transduction system, GT – glycosyltransferase, poly exp – polysaccharide export, pp – periplasmic, put omem – putative outer membrane.

The orfC1-4-31 has a maximum coding capacity of 136 amino acid residues. The first two codons are for arginine. Therefore, the peptide could only be 134-residues long. The ORF orfC1-4-31 is highly homologous to GenBank domain smart00448 (REC), a CheY-like receiver domain (Score101 E 5e-23). CheY regulates the clockwise rotation of *E. coli* flagellar motors. This domain contains a phosphoacceptor site that is phosphorylated by histidine kinase homologues. Signal transduction genes have long been known to be involved in *M. xanthus* development. Among the five *M. xanthus* developmental signals, A-signal transduction is known to use elements of the two-component signal transduction systems, such as *asgA*¹ and *asgD*. The *sasA* and *sasR* genes are another pair of the two-component system elements involved in the regulation of A-signal-dependent gene Ω 4521. Another example of a signal transduction component involved in *M. xanthus* development is the *frzE* gene, which is homologous to chemotaxis genes cheA and cheY. The ORF orfC1-4-31 has high homology to signal transduction histidine kinase (a sensor of two-component signal transduction system) from *Nostoc punctiforme* and *Thermosynechococcus elongates*. Since orfC1-4-31 is downstream from the *cds29*, *cds 32*, and *cds33*, the defects observed in these mutants could be due to their polar affects on orfC1-4-31. It is important to note that there are more than 50 two-component signal transduction coding fragments (E value < e⁻³⁰) in the *M. xanthus* genome. Within the annotated regions of the genome in this dissertation, it is quite clear that the two component system elements are spread into many operons, particularly into polysaccharide synthesis operons.

orfC1-5-24 is homologous (score 47.8, E 3e⁻⁰⁶) to the conserved domain COG3206 GumC, which is involved in exopolysaccharide export in the periplasm in *Xanthomonas campestris*

¹ Refer to Chapter 3 for discussion and references.

(Vojnov *et al.*, 2003). However, it did not match any specific database sequence. This is probably why this kind of conserved domain databases is useful in guiding research. It at least points out the possible direction for further investigation. This ORF was mutagenized using internal fragment replacement mutagenesis, and the results are presented in the section for the Internal Fragment Replacement Mutagenesis Analysis. The results suggest the involvement of orfC1-5-24 in development.

ORF orfC1-5-23 is highly homologous to three conserved domains: (1) COG1215 glycosyltransferases, probably involved in cell wall biogenesis; (2) COG0463 WcaA, glycosyltransferases involved in cell wall biogenesis. COG1215 and COG0463 seem to represent the same functional domain conserved among glycosyltransferases involved in cell wall biosynthesis, such as the protein WcaA in *E. coli* and *Mycoplasma gallisepticum*; (3) pfam00535 glycos_transf_2, a diverse glycosyltransferase family, transferring sugar from UDP-glucose, UDP-N-acetyl-galactosamine, GDP-mannose or CDP-abequose to a range of substrates including cellulose, dolichol phosphate and teichoic acids. However, orfC1-5-23 found highest homology with some uncharacterized genes: accession numbers ZP_00018916 from *Chloroflexus aurantiacus* (Score 188, $E\ 2e^{-46}$) and ZP_00129998 from *Desulfovibrio desulfuricans* (Score 162, $E\ 2e^{-38}$). The exopolysaccharide transferase component gene *epsO* from *Methylobacillus sp.* 12S (Yoshida *et al.*, 2003) has a very high homology to orfC1-5-23 at Score 148, and $E\ 2e^{-34}$. Therefore the most likely conclusion is that orfC1-5-23 codes a *wcaA* like gene involved in transferring glycosyl moieties for cell wall biosynthesis.

The last ORF on this transcript is orfC1-4-29. Although its G+C bias in the third base of the codon is only 75%, it carries a conserved domain, COG5340 (Score 72.4, $E\ 4e^{-14}$), predicted to be a transcriptional regulator (Figure 1.12). It also has a highly homologous sequence from *Mycobacterium tuberculosis* with a Score of 106 and the E value of $2e^{-22}$. However, this database sequence is labeled as a hypothetical protein RV1044. There is no experimental evidence for its functions.



Figure 1.12 Alignment between the orfC1-4-29 and the transcriptional regulator domain COG5340 members. Identical residues are highlighted in red, aligned regions are in blue, and unaligned in grey.

Another interesting ORF in this region is the orfC1-2-19+3-29. This ORF is coded in the opposite direction, compared with the others in the neighborhood, and leaves a big gap (~2 kbp) between itself and the ORFs orfC1-4-29. It consists of two small ORFs with a 144-base overlap. They are both homologous to the same database sequences, a transposase active in *Bradyrhizobium* (score 107, E $6e^{-45}$ and score 98.2, E $6e^{-45}$, respectively). One of the two codes for the carboxyl terminal part, while the other codes for the amino terminal part of the same transposase. However the two ORFs have a 48-codon overlap. It is known that this transposase is inactivated in some hosts by truncation. Therefore it is more likely that these two ORFs suggests that the transposase have gone through some inactivation process in the *M. xanthus* genome. These two transposase ORFs lie in a region where the third base in the codons are only weakly biased for G+C. Although the predicted transposase ORFs do not follow the general tendency for *M. xanthus* codons, their nature of being transposase argues strongly for their being authentic ORFs. This could mean that the transposase is a foreign gene that gained access to *M. xanthus* in the past. It would be very interesting to test the gene's functions, to see whether they are active or whether they can be activated. A BLAST search in the *M. xanthus* genome indicated a few other loci harbouring the same or similar transposases. Potentially these transposases could be genetic and molecular biological tools. Of immediate interest though is that these two ORFs and a big gap between the transposase and the orfC1-4-29 are strong evidence for the end of the apparent transcript discussed so far.

A few ORFs between the *cds19* operon and this long one are worth noting. The orfC1-5-19

and orfC1-6-26 are highly homologous. The amino terminal fragment (152 amino acid residues) of orfC1-5-19 is highly homologous (Score 209, $E\ 1e^{-52}$) to the gene *oar*, an ompA-related gene known in *M. xanthus* (Matinez-Canamero *et al.*, 1993). Since gene *oar* is required for *M. xanthus* development, it is likely that orfC1-5-19 is developmentally regulated as well. The rest of the ORF has high homology (Score 171 $E\ 3e^{-41}$) to a von Willebrand factor type A protein – a putative outer membrane protein or exported protein. In mammals, the glycoprotein von Willebrand factor (VWF) is an adhesive protein involved in hemostasis.

Interestingly, the orfC1-6-26 also matches the von Willebrand factor. A close inspection revealed that the two potential ORFs have an overlap of seven codons, and match two consecutive regions of the database VWF sequences. This suggests strongly that the *M. xanthus* genomic sequence in the GenBank contains an error in or near the overlapping 7 codons, causing a frame shift and a stop codon downstream. A careful examination¹ of the sequence found that the ORF truncation was probably due to two separate extra G/C base insertions in the middle of the protein. Once the insertions are removed, and the orfC1-6-26 and orfC1-5-19 are joined, the new ORF has 720 codons, initiation codon at codon 21ATG. This protein is 700-residues long, with a predicted signal peptide of 22-amino acids long (Bendtsen *et al.*, 2004). It has a good homology to VWF domain COG2304 (Score 107, $E\ 4e^{-24}$, an improvement from Score 80.7, $E\ 3e^{-16}$). Its best homologous protein is the von

¹ A BLASTX search showed that the pieces of sequence matched the VWF database sequences tripped from frame -3 to -1 and then -2. The obvious correction is to resplice the matching sequence into a single frame. There are many options: the one used is to make a minimal deletion. The two bases removed are the Gs after codons 473TGG and before 510GGC in the resultant sequence. The original separated ORFs have homology values at Score 209 $E\ 1e^{-52}$ and Score 139 $E\ 5e^{-32}$. The corrected sequence has an overall Score 323 $E\ 1e^{-86}$ to *shewanella oneidensis* MR-1 protein NP_719099.11 [accession numbe], a von Willebrand factor type A domain protein. Judging based on the improvement of the resultant homology, the sequence revision is probably correct.

Willebrand factor type A domain protein (access number ZP_719099.11) from *Shewanella oneidensis* MR-1 (Score 323, E $1e^{-86}$).

VWF exists in many prokaryotes, such as *Pseudomonas fluorescens*, *Clostridium thermocellum*, and *Escherichia coli*. In mammals, plasma von Willebrand factor (VWF) is a multimeric glycoprotein from endothelial cells and platelets that mediates adhesion of platelets to sites of vascular injury. VWF levels are associated with markers of increased oxidative stress and therefore reflect the severity of biochemical abnormalities, which contribute to diabetic vascular disease (Ibrahim *et al.* 2004; Ruggeri 2003). Supporting with the predicted signal peptide, a domain pfam05738 was found in the amino half of the protein. The pfam05738 is a B-type domain in Cna protein. This domain is found in *Staphylococcus aureus* collagen-binding surface protein. However, this region does not mediate collagen binding, instead it forms a beta sandwich structure. It is thought that this region forms a stalk in *Staphylococcus aureus* collagen-binding protein that presents the ligand-binding domain away from the bacterial cell surface. This could mean that the VWF domain in *M. xanthus* may actually be a cell surface protein and mediate some kind of adhesive function.

Cluster 2 (Figure 1.13)

The Cluster 2 is at the 3.8 Mbp position on the *M. xanthus* physical map (Figure 1.10). The insertion *cds4* is mapped to Cluster 2 at base position 3910. It is in a known *M. xanthus* gene *tgl*¹. Protein Tgl is believed to be a factor required for the assembly of previously synthesized pilin subunits (Wall *et al.*, 1998). Since without Tgl *M. xanthus* cells cannot assemble pilins into pili (Rodriguez-Soto and Kaiser, 1997), one would assume the *cds4* cells to be completely deficient in social motility. However, we found the *cds4* mutant still has some remnant social motility on low percentage agar (Figure 1.4, Table 1.2). Surprisingly, the Tgl protein has a good homology (Score 62.3, $E\ 8e^{-11}$) over a 173-residue region to the domain COG3063, characteristic of PilF. Therefore, the stimutable gliding motility is probably due to the function of PilF. This will make the Tgl protein more identifiable/comparable with other known proteins. PilF is known to function in export or assembly of fimbriae/pili in *Pseudomonas aeruginosa*. The ORF upstream from the Tgl is called ORFA (Rodriguez-Soto and Kaiser, 1997), with no known functions and no homolog in the GenBank. Upstream from the ORFA, there is an ORF-like region with extremely high GC bias (overall G+C 85%), probably indicating the upstream limit for this apparent operon. The ORF downstream from *tgl* has homology to a bacterial DNA-binding motif, found in a transcriptional regulator from *Geobacter sulfurreducens*. Further downstream is a RecO domain (Score 134, $E\ 2e^{32}$) for recombinational DNA repair protein such as *recO* from *Geobacter sulfurreducens* (Score 128, $E\ 1e^{-28}$). This operon ends with a fructokinase PfkB similar to the one from *Bacillus halodurans* (Score 195, $E\ 1e^{-48}$).

¹ Comparing to the *M. xanthus* genome database sequence, we noticed a base deletion (C686, count from the start codon) in the sequence submitted by Rodriguez-Soto and Kaiser (1997) to the GenBank. This leads to frame shifted protein codes at the carboxyl terminal and an early termination codon in their predicted protein sequence. Our predicted protein sequence is 253 amino acid residues long, while the previous prediction is 241 amino acid residues long (Rodriguez-Soto and Kaiser, 1997).

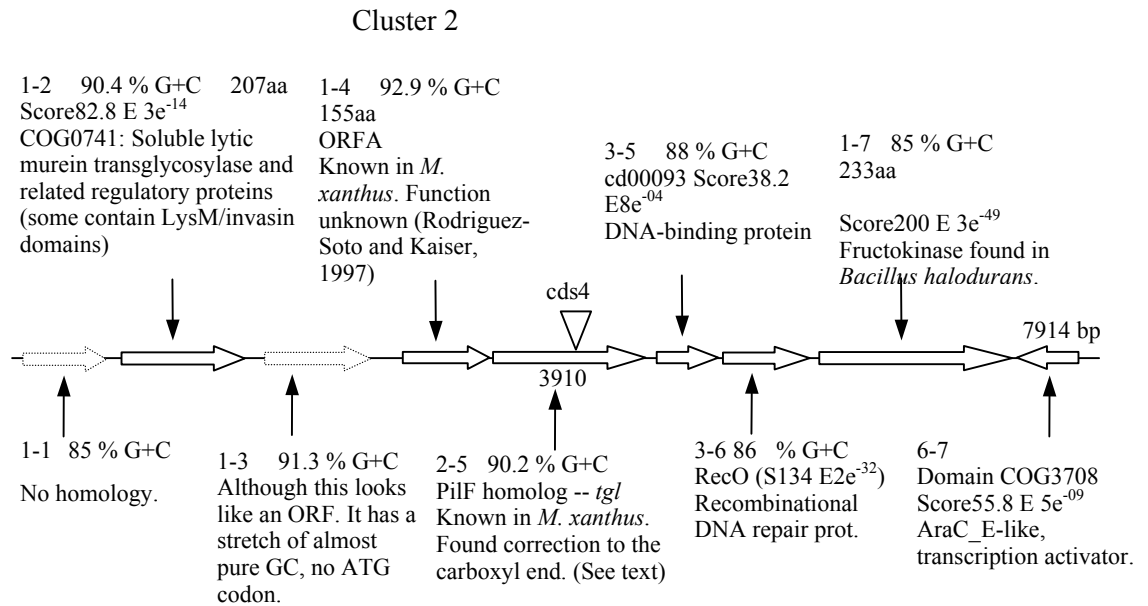


Figure 1.13 Map of insertion *cds4*.

The *pilF* (*tgl*) gene is the only pili-related gene that is not encoded in the *pil* cluster of genes. PilF (Tgl) protein is required for pili formation, yet can be provided externally from other cells (Wall *et al.*, 1998). Recent experiments show that PilF (Tgl) is an outer membrane protein (Simunovic *et al.*, 2003). The *cds4* results demonstrate that the *pilF* (*tgl*) gene is required for polysaccharide biosynthesis, probably involved in forming a functional exopolysaccharide (type IV pili) transporter system (Figure 1.24). This result is consistent with and complementary to previous observations (Wall *et al.*, 1998; Simunovic *et al.*, 2003).

Cluster 3 (Figure 1.14)

Insertion *cds1* is at Cluster 3 (5.4 Mbp on the *M. xanthus* map [Figure 1.10]) in an ORF (orfC3-3-2) that has no homolog in the GenBank. But it does have other qualities for an authentic open reading frame. It contains a single methionine, codon 18. It has high G+C bias at 92.9% on the third base of its codons, potential Shine-Dalgarno sequence (GAAGG) upstream from the predicted start codon, and 377 amino acid residues long. It seems to be a cytoplasmic protein since it does not have a signal peptide (Bendtsen *et al.*, 2004). The protein sequence has good homology to two hypothetical GenBank sequences: protein GSU1932 (accession NP_952980.1) from *Geobacter sulfurreducens* PCA (Score 103, $E 5e^{-21}$) and protein Gmet02000005 (accession ZP_00301574.1) from *Geobacter metallireducens* GS-15 (Score 97.1, $E 7e^{-19}$).

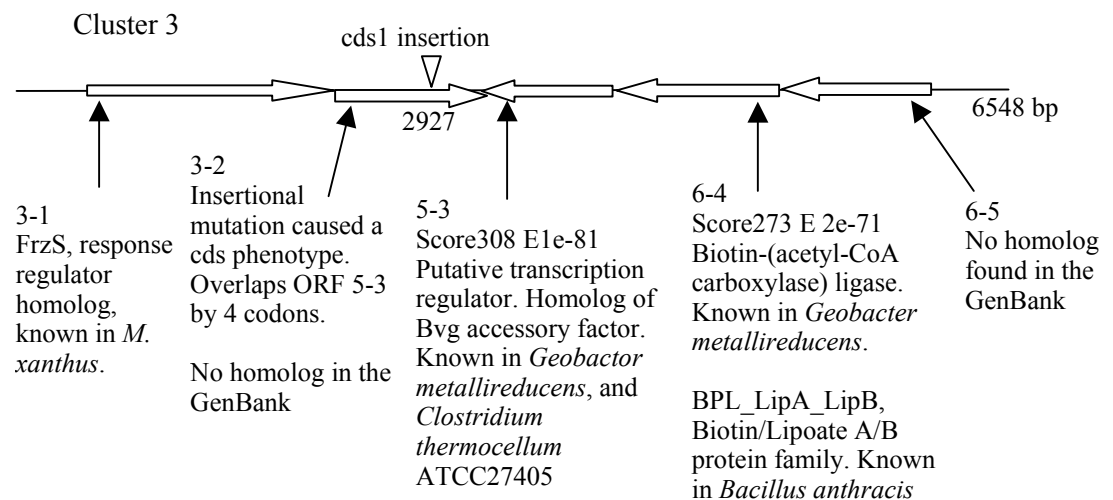


Figure 1.14 Insertion map of *cds1* at Cluster 3. Notice the overlap of 13 bases between orf1-3-2 and orf1-5-3.

The insertion is in the downstream one third of the predicted ORF orfC3-3-2. This mutant displays smooth colony morphology, much reduced Calcofluor White-binding capacity (Figs.

1.7 and 1.8), and dramatically reduced cohesion efficiency (Fig. 1.4), yet retains a considerable amount of S-motility (Figure 1.5, Table 1.2). Since the insertion *cds1* is at the end of a apparent operon, there are probably no polar effects on any other genes. The observed phenotype must be caused by the insertion mutation. This fact authenticates the *orfC3-3-2* as a real and functional gene, presumably required for exopolysaccharides production, cohesion, and development (Figures 1.5, 1.6, 1.7, 1.8). It has a mild affect on S-motility (Table 1.2).

Cluster 4 (Figure 1.15)

Insertion cds13 is the only insertion at Cluster 4 (map position 6.11 Mbp). It is in a potential ORF orfC4-1-2. The orfC4-1-2 has a predicted length of 684 amino acids, and a G+C content of 85%. The amino terminus is homologous to a conserved domain KOG4300, for methyltransferase (Score 51.9 and E of $5e-07$) and COG2227 from protein UbiG, 2-polyprenyl-3-methyl-5-hydroxy-6-methoxy-1,4-benzoquinol methylase (Score 49.5 E $2e^{-06}$). The closest database homolog (score 54 and E $7e-06$) is the SAM-dependent methyltransferase from *Ralstonia metallidurans*. The carboxyl two thirds of the ORF has no homolog at all. The insertion cds13 does not seem to severely affect S-motility and development. However, its Calcofluor White-binding capacity is dramatically reduced (Figures 1.7 and 1.8), probably due to the polar effect on the downstream ORF orfC4-3-4 which is a glycosyltransferase (Yang *et al.*, 2000).

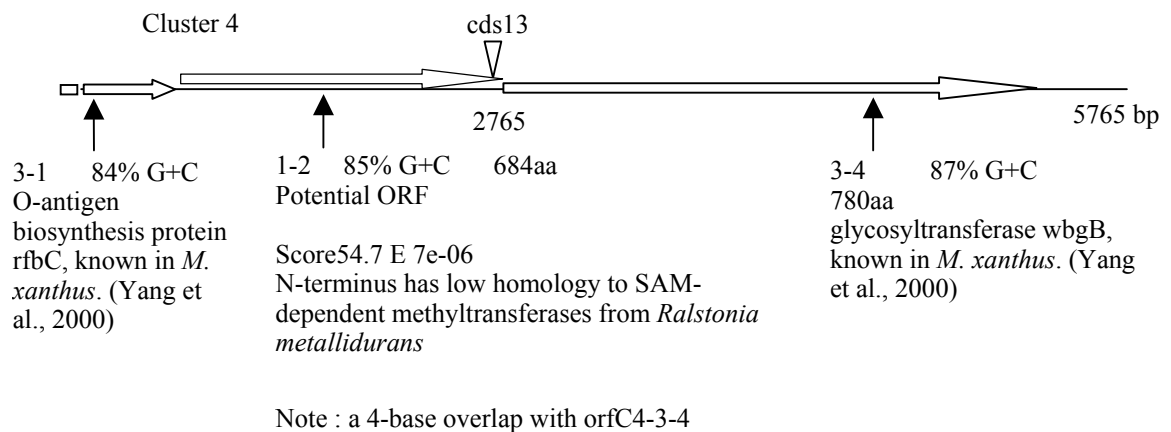


Figure 1.15 Map of insertion cds13 at Cluster 4.

Cluster 5 (Figure 1.16)

One cluster that had been most heavily hit by the *magellan-4* transposon is the well-studied pilus biosynthesis gene cluster at Cluster 5 (map position 7.5 Mbp). Among the unique 17 *magellan-4* insertions, 9 are in the 22 kb region of the *pil* gene cluster. From upstream to downstream in the transcript are *ribF*, *pilBTCSRAGHID*, then a signal transduction sensor-activator pair *pilS₂R₂* (Wall and Kaiser, 1999), followed by *pilMNOPQ*. A short expression is *pilBTCSRAGHIDS₂R₂MNOPQ*. However, the apparent operon is even longer: it is over 38 kb, because at the end of the *pil* gene cluster there are ten additional ORFs tightly coded in the same direction (Fig. 1.24). This makes the apparent operon size extraordinarily long, but there is no easy way to computationally define it any better at this time. This is the first time for such a complete *pil* gene cluster being described in *M. xanthus*. In the *pil* gene cluster, the *ribF* and *pilMNOP* genes have never been described before. We have two insertions *cds22* and *cds42* in the *pilO* gene. Both of these insertion mutants are deficient in social motility and Calcofluor White-binding and development (Figures 1.3, 1.4, 1.5, 1.6, 1.7, and 1.8). The deficiency of strain *cds42* seems slightly less severe than *cds22*.

The first ORF on the *pil* transcript is *orfC5-3-35*, which has a capacity of 503 amino acid residues. But the first two fifths portion of the ORF has an average third base G+C bias of 70%, and the first methionine did not appear until position 196. After the first methionine, the G+C bias reached 87%, and the sequence matched the COG0196 (score 216 E1e-56), a RibF domain, which is involved in FAD synthesis. Its closest homology is found with FAD synthase from *Thermoanaerobacter tengcongensis* (159/1e-37).

It should be noted here that Wu and Kaiser deposited this region of the sequence from

Myxococcus xanthus in the GenBank in 1997. However, they were not able to predict any function for this piece of sequence at that time. I found that their sequence has significant differences from the one I retrieved from the NCBI incomplete genome database. Now I predict with good confidence that the function for this ORF is riboflavin synthesis.

Insertions cds14, cds20, and cds37 are the same insertion but collected multiple times. This insertion will be referred to as cds14 in the following discussion. Insertion cds14 was located in the beginning of the *pilB* gene. The *pilB* gene is very highly homologous to a domain COG2804 of PulE protein (Score 552, $E7e^{-158}$). PulE is a component of ATPase involved in Type II secretory pathway and pilus assembly pathway. Its homology with type IV pilus biogenesis protein PilB *Geobacter sulfurreducens* PCA is very high too (Score 742, E 0.0). The gene *pilB* and *pilC* are related to the membrane traffic ATPase and the inner membrane proteins of type II secretory systems. *pilT* encodes an ATPase that is responsible for pilus retraction in *M. xanthus*, providing motility force, but not piliation (Wu et al., 1997). Deletion mutation analysis shows that *pilB* and *pilC* are required for pilus biogenesis (Wu et al., 1997)

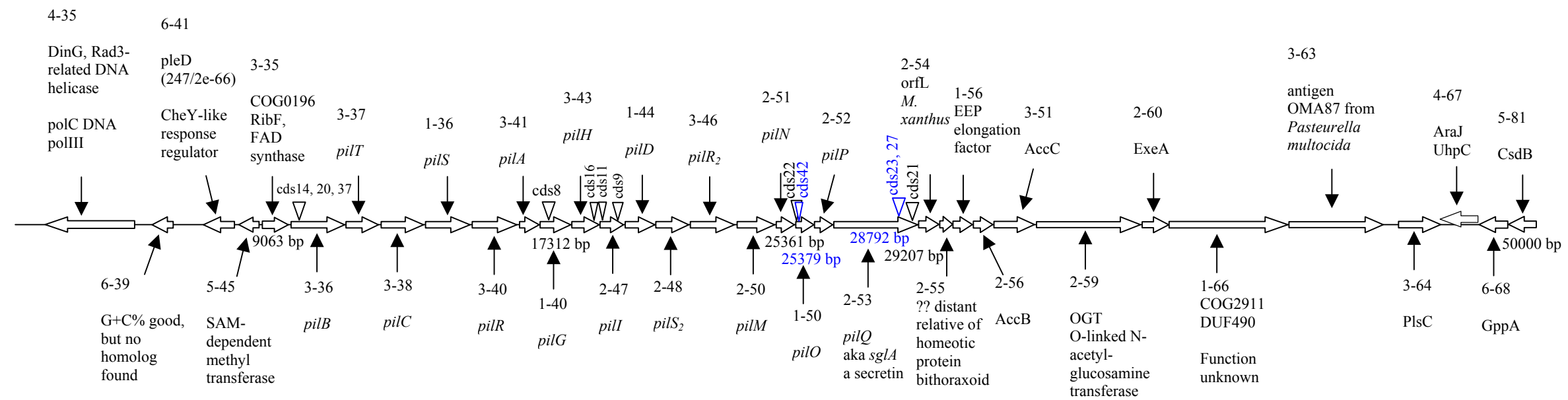


Figure 1.16 Transposon insertion sites in the Cluster 5. Nine insertions are located in this single pil gene cluster. There is one insertion in each of *pilB*, *pilG*, *pilH*. There two insertions in each of *pilI*, *pilO*, and *pilQ*.

The insertions *cds8*, *cds9*, *cds11* and *cds16* should be considered together because they are in three genes *pilGHI*, that code for the ABC-type complex required for type IV pilus biogenesis and social gliding motility in *Myxococcus xanthus*. The complex may also participate in pilus assembly and/or the export of the pilin PilA protein (Wu et al., 1998). The complex is a multidrug transport system in many organisms, *Bacillus halodurans*, *Encephalitozoon cuniculi*, *Borrelia burgdorferi* (NCBI database). Insertion *cds8* is in the *pilG* gene. Insertion *cds16* is in the *pilH* gene. Insertions *cds11* and *cds9* are in the *pilI* gene. These four mutants produce the same phenotype in growth (Fig. 1.3), motility (Fig. 1.5; Table 1.2), development (Fig. 1.6), and Calcofluor White-binding (Figs. 1.7 and 1.8). In cohesion assays the strain *cds9* seems to perform slightly better than the other three (Fig. 1.4).

Here again, a pair of genes with high homology to 2-component signal transduction system sensor and activators is inserted in an otherwise pure *pil* gene cluster. The *orfC5-2-48* and *orfC5-3-46* have not been described in the published literature, and are not characterized any further here. Nevertheless, a repeated observation is that *M. xanthus* tends to intercalate components of 2-component signal transduction systems into polysaccharide production and pilus biogenesis operons.

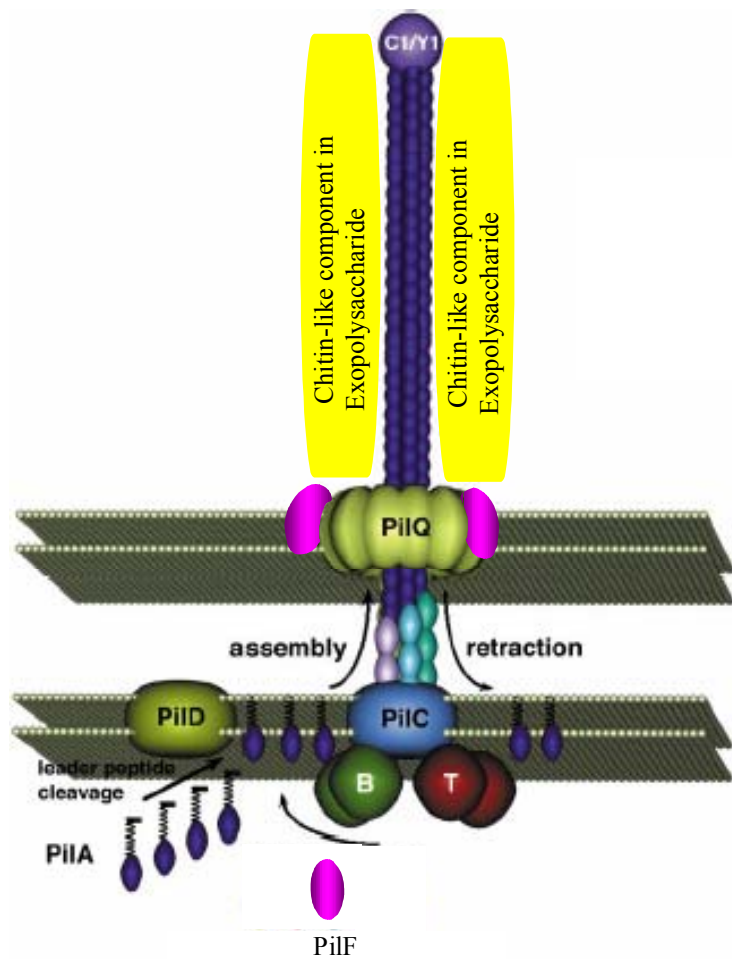


Figure 1.17 A model type IV pilus system. Adapted from Mattick, 2002, showing the unique component PilF in the outer membrane position, and the chitin-like element in the exopolysaccharide required for pilus retraction.

Insertions *cds42* and *cds22* both are in a potential ORF that is highly homologous to the *pilO* gene. The *pilO* gene, together with its neighbors, a predicted *pilN* located 5' of this gene, and a *pilP* downstream, are all involved in pilus assembly. In *Pseudomonas aeruginosa* *pilO* together with *pilN*, and *pilP* forms a pilus assembling complex. In *Thermus thermophilus* HB27, pili and natural competence are linked via these *pil* genes and others. The phenotypes of these strains are very similar. They all have severe defects in growth (stationary phase, Fig. 1.3), motility (Fig. 1.5; Table 1.2), development (Fig. 1.6, pages 25 and 26), and Calcofluor White-binding (Figs. 1.7 and 1.8). Strain *cds22* has better cohesion ability than *cds42*, this is also somewhat reflected in the development. The strain *cds22* has a low degree of aggregation on the starvation agar, whereas *cds42* is completely unable to aggregate.

Insertions *cds21*, *cds23*, and *cds27* are located in the known *M. xanthus* gene *pilQ*. *pilQ* is a secretin, forming a ring structure of 10 to 18 secretin subunits (Figure 1.17). These cylindrical structures, as visualized by electron microscopy, have central cavities ranging in size from 50 to 95 Å, and are involved in macromolecular transport across the outer membrane. Such cavities are large enough to accommodate the transportation of folded proteins and assembled macromolecular complexes, such as filamentous phage (diameter, 65 Å) or type IV pili (diameter, 2 Å) (Wall *et al.*, 1999).

Insertion *cds21* is near the carboxyl end of the *pilQ* gene, while insertions *cds23* and *cds27* are the same insertion and are located 400 bp upstream of *cds21*. The *pilQ* gene is also known as *sglA*, which is also the gene in which the strain FB and DZ101 has base change mutations. All these three insertions caused developmental defects, while FB and DZ101

develop normally. This signifies that strain FB and DZ101's ability to develop normally is a special case in the vast amount of strains that display smooth colony morphologies. Since it is known that S-motility is required for development, and that chitin is required for the pilus retraction and for generating S-motility force (Li *et al.*, 2003), the base changes in the *pilQ1* allele of strain FB and the like must be such that under the development conditions enough chitin is released via PilQ1 transporter to support a certain amount of S-motility that is required for what we consider as the normal development. Obviously, it is impossible for *pilQ*-insertion mutants to behave this way.

Among all those insertion mutants in this *pil* gene cluster, only the *pilQ* is obviously related to the loss of polysaccharide production. Since *pilQ* is a secretin, loss of *pilQ* could have hindered the export of polysaccharide to the outer surface of the cell, preventing the fibril formation. Since N, O, P are believed to be involved in pilin assembly, it is quite puzzling: what should pilin production and pilus assembly have to do with polysaccharide production? Why should mutants cds8, 9, 11, 16, 22, and 42 be defective in polysaccharide production? The transporter for polysaccharide PilQ is supposed to be intact in all these cases. The most probable explanation is that the observed defects are due to the polar effects of the insertions on the downstream genes in the transcript. Since the *pilQ* gene is at the end of the whole *pil* cluster, all the defects may demonstrate the fact that this *pil* gene cluster indeed is a gigantic operon, and all the similarities in each and every insertion mutant simply shows the fact that they have the same polar effect on the *pilQ* gene. In other words the phenotypes observed in the insertion mutants in this region could be all due to the loss of the gene product from *pilQ*.

Other ORFs on the apparent *pil* transcript

The long ORF orfC5-2-59 has a homology for the N-terminal region to KOG4626 (score 55.4, $E = 8e^{-08}$), from O-linked N-acetylglucosamine transferase OGT involved in carbohydrate transport and metabolism, posttranslational modification, protein turnover, chaperones, and signal transduction mechanisms. Defects or polar effects on this gene could be quite severe.

The ORF orfC5-2-60 has strong homology with ExeA (score 137, $E = 3e^{-33}$), which is a component of type II secretory pathway, possibly an ATPase, involved in intracellular trafficking and secretion.

The long ORF orfC5-1-66 is homologous to an unknown bacterial Conserved Domain COG2911 and DUF490 (score 100, $E = 3e^{-21}$). But these domains have no predicted functions at this time.

Another long ORF orfC5-3-63 is homologous to COG4775, characteristic of outer membrane protein or putative (protective) surface antigen OMA87 (score 172, $E = 4e^{-43}$). This family includes the following surface antigens: D15 antigen from *Haemophilus influenzae*, OMA87 from *Pasteurella multocida*, OMP85 from *Neisseria gonorrhoeae*. The family also includes a number of eukaryotic proteins as well that are members of the UPF0140 family.

The ORF orfC5-3-64 is homologous to COG0204 (score 78.4, $E = 3e^{-15}$), from protein PlsC 1-acyl-sn-glycerol-3-phosphate acyltransferase, which functions in phospholipid biosynthesis and has glycerolphosphate, 1-acylglycerolphosphate, or 2-acylglycerolphosphoethanolamine

acyltransferase activities. Tafazzin, the product of the mutated gene in patients with Barth syndrome, is a member of this family.

After this long chain of apparent *pil* operon genes comes the first ORF transcribed in reverse direction orfC5-4-67. The orfC5-4-67 is homologous to domain COG1330 (score 190, $E\ 8e^{-49}$), from the family of the sugar transporter (spinster) transmembrane protein, AraJ, engaged in carbohydrate transport and metabolism.

The ORF orfC5-6-68 is homologous to domain COG0248 (score 221, $E\ 5e^{-58}$) from protein GppA, an exopolyphosphatase involved in nucleotide transport and metabolism and in inorganic ion transport and metabolism.

The ORF orfC5-5-81 is homologous to several domains all in the amino acid transport and metabolism pathway. (1) COG0520 (score 185, $E\ 1e^{-47}$) is from the CsdB family, a selenocysteine lyase. (2) KOG1549 (score 157, $E\ 4e^{-39}$) is from the cysteine desulfurase NFS1 family. (3) COG1104 (score 83.3, $E\ 1e^{-16}$) is from the NifS family, a cysteine sulfinatase desulfurase or cysteine desulfurase and related enzymes. However, its highest protein homology (score 359, $E\ 6e^{-98}$) is the protein isopenicillin N epimerase from *Ralstonia solanacearum*, which is a key enzyme in committing the β -lactam antibiotic biosynthesis intermediate isopenicillin N to form a more potent endproduct, cephalosporins. It not clear whether *M. xanthus* has the full complement of genes necessary to make cephalosporins. However, it is well known *M. xanthus* produces antibiotics and is naturally resistant to ampicillin (a member of the penicillin family) and many other antibiotics.

The first ORF transcribed away from the *pil* operon is orfC5-5-45, which has strong homology with conserved domain COG0220 from S-adenosylmethionine-dependent methyltransferase (score 128, E $2e^{-30}$). Its closest protein homolog is from *Geobacter sulfurreducens* (score 107, E $4e^{-22}$).

Following in the same direction is orfC5-6-41, which has very strong homology to the CheY-like regulator domain COG3706 (score 247, E $2e^{-66}$) from the PleD family. Its closest protein is found in the protein ZP_00081495 from *Geobacter metallireducens* (score 209, E $6e^{-53}$).

The ORF orfC5-4-35 is homologous to COG1199 (score 228, E $7e^{-60}$) from the DinG family, a Rad3-related DNA helicase involved in transcription, DNA replication, recombination and repair.

Cluster 6

Cluster 6 is at the 8.6 Mbp position on the *M. xanthus* physical map (Figure 1.10). Insertion 24 and 40 are the same. They will be referred to as cds24. Insertion cds24 is in the carboxyl end of the *sglK* gene. The SglK protein is a chaperone protein, belonging to the dnaK class. However experimental evidence showed that *sglK* does not respond to heat shock (Weimer *et al.*, 1998). This mutant forms smooth colonies, has lost most of its Calcofluor White-binding capacity, is unable to form aggregates or sporulate on development agar. It is interesting to note that a transposon insertion in another *dnaK*¹ gene, *stk*, causes *M. xanthus* cells to form colonies that appear drier and rougher than wild type DK1622, and produce an extra amount of polysaccharide (Kim *et al.*, 1999). Although *sglK* is known to be essential for S-motility and development (Yang *et al.*, 1998; Weimer *et al.*, 1998), this is the first link of the *sglK* gene to exopolysaccharide production.

Insertions cds18 and cds28 are in a known *M. xanthus* gene, *difE* (Lancero *et al.*, 2002). The *dif* genes are known as the second chemotaxis system in *M. xanthus*. DifE protein is a CheA-like kinase with a strong homology to CheA domain COG0643 (score 313, E 1e⁻⁸⁵). The gene *difE* is in the middle of a long stretch of annotated *M. xanthus* genome that is known to be important for motility and development (Lancero *et al.*, 2002).

I noticed that there are two tandem *pilT* genes in this region, about 14 kb downstream from the insertion cds28. This makes a total of three *pilT* genes described in this dissertation. A BLAST search of the incomplete *M. xanthus* genomic sequence database at NCBI using the

¹ There are at least 10 *dnaK* homologs in the *M. xanthus* genome with E less than e⁻³⁰ when the genome is searched using BLAST with the *dnaK* consensus sequence as the query sequence.

pilT consensus found a total of seven *pilT* homologs in the *M. xanthus* genome. They are distributed over four contigs: One on the contig526, two on contig521, three on contig503, and one on the contig520. The one on contig526 is the one described in the huge *pil* gene cluster above. Two of the three on the contig503 are at Cluster 6. Among the other four *pilT* homologs not described in the six polysaccharide production clusters, at least one of them has better homology than the two in Cluster 6. That is, the one on the contig521 has better homology with *pilT* consensus sequence than the two at Cluster 6 on contig503, with a 52% identity, and 70% positive over 351 residues.

Table 1.6 The distribution of the *pilT* -related genes in the *Myxococcus xanthus* genome

Contig #	Base positions	Score and E value	Identity (%), similarity (%), and number of amino acid residues
526	147727-148863	370 / e-104	53%, 74%, 347
521	710227-709166	352 / 3e-98	52%, 70%, 351
521	116599-117399	134 / 8e-33	34%, 53%, 267
503	935385-936386	349 / 2e-97	54%, 72%, 334
503	934184-935230	327 / 6e-91	48%, 68%, 350
503	1447422-1448210	127 / 2e-30	36%, 53%, 223
520	62240-63355	133 / 2e-32	31%, 46%, 331

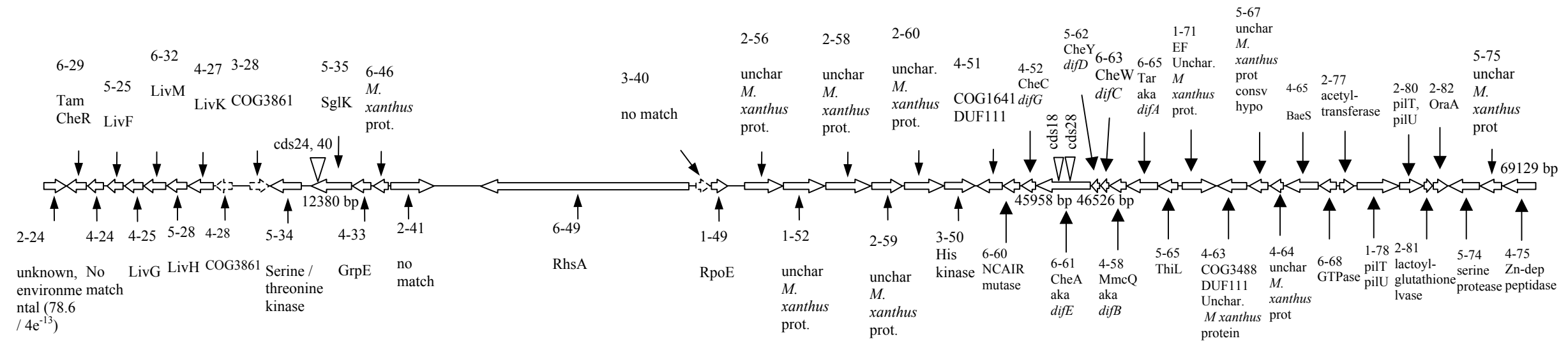


Figure 1.18 Transposon insertion sites in the Cluster 6. Nine insertions are located in this single pil gene cluster. There is one insertion in each of *pilB*, *pilG*, *pilH*. There two insertions in each of *pilI*, *pilO*, and *pilQ*.

Internal Fragment Replacement Mutagenesis Analysis of Two Selected Operons

Exopolysaccharide biosynthesis is a complex process. It involves many enzymes. Consequently, glycosyltransferases are a diverse family. They function in a variety of different steps of the exopolysaccharide biosynthesis: basic sugar synthesis, polymerization, transportation to the outer membrane or outside of the cell, etc. Many glycosyltransferases are found in this study. Some ORFs have no homology to any database sequences. Since in bacteria genes involved in similar functions are often organized in operons, we assume those ORFs, particularly those with no database homolog, in the same apparent operons as cds insertions are likely to be involved in the production of exopolysaccharide. A number of ORFs were selected for mutagenesis analysis, marked with red arrows at Cluster 1 (Figures 1.11, 1.19, and 1.20). Three ORFs upstream from insertion SR53, i.e. outside the operon SR53, and three from within the cds29 operon, i.e. downstream of cds29 insertion were mutated with Internal Fragment Replacement Mutagenesis method as described in the Materials and Methods section. These six mutations are also marked as 53-3, 53-4, 53-6 for those near SR53, and 29-3, 29-5, and 29-6 for those near cds29.

The mutation 53-3 is in orfC1-6-2, which has strong homology to protein EpsP (score 176, E value $6e^{-43}$), a component of a glycosyltransferase system (EpsBJNOP and EpsR) involved in the synthesis of the repeating unit of methanolan (one kind of exopolysaccharide, composed of glucose, mannose and galactose) onto the lipid carrier in *Methylobacillus sp.* 12S (Yoshida *et al.*, 2003). Besides, orfC1-6-2 has a conserved domain homologous to (1) COG1922, WecG, Teichoic acid biosynthesis proteins (score 207, E value $2e^{-54}$); (2) pfam03808, Glyco_tran_WecB, Glycosyl-transferase WecB/TagA/CpsF family (score 168, E value $1e^{-42}$). This strongly suggests that the orfC1-

6-2 encodes a component of a glycosyltransferase complex. We created a truncation mutation 53-3 in this ORF to test the effect of loss of this protein.

The orfC1-5-4 has a strong homology to endoglucanase A precursors (endo-1,4-beta-glucanase) from *Bacillus lautus* (Score 170, $E 5e^{-41}$) and *Clostridium acetobutylicum* (score 148, $E 2e^{-34}$). Endoglucanase A is also called cellulase A, and EG-A in *Bacillus lautus*. This ORF was mutated with the internal fragment replacement mutagenesis method (see Materials and Methods section for details) for testing its phenotype.

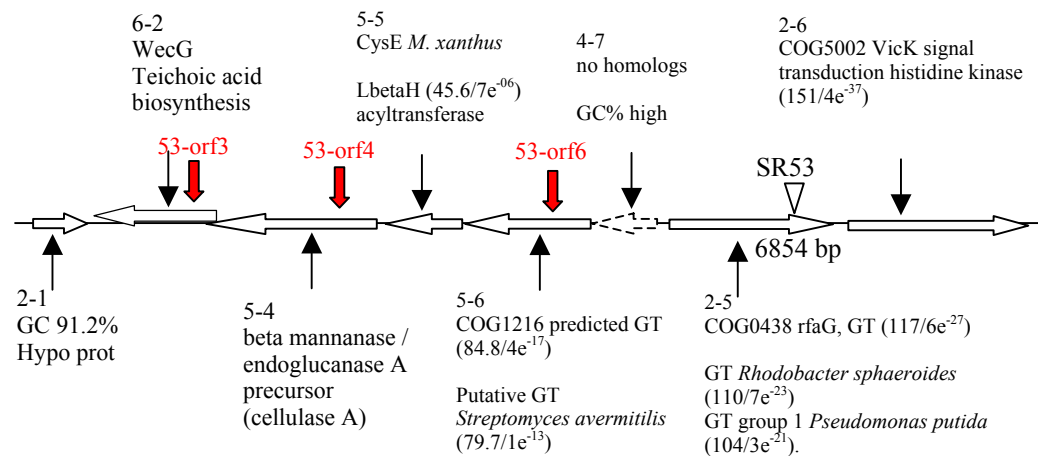


Fig. 1.19 Internal fragment replacement mutagenesis sites near SR53 at Cluster 1. Red arrows indicate the site of targeted mutations.

The orfC1-5-5 has good homology (score 106, $E 2e^{-27}$) to serine acetyltransferase from *Clostridium acetobutylicum*. It hits many conserved domains. Two examples are: (1) COG1045, CysE, serine acetyltransferase (Score 114, $E 7e^{-27}$), involved in amino acid metabolism. (2) COG0110, WbbJ, acetyltransferase (isoleucine patch superfamily) (score 72.6, $E 2e^{-14}$). All the homologies indicate that the orfC1-5-5 encodes an acetyltransferase. The first 96 amino acid residue sequence of this gene from *M. xanthus* has been deposited in

the GenBank but there is no description or literature on it. Here the full length of the sequence is used in this dissertation, and it supports the earlier identification of this gene.

The highest homology of orfC1-5-6 is found with a hypothetical protein from *Chloroflexus aurantiacus*. But it also has a good homology to a putative glycosyltransferase known in *Streptomyces avermitilis* MA-4680 (score 79.7 E $8e^{-14}$). When searched for domain homologies, orfC1-5-6 was found homologous to several conserved domains: (1) COG1216, a predicted glycosyltransferases (score 84.8, E $2e^{-17}$); (2) COG1215, glycosyltransferases, probably involved in cell wall biogenesis; (3) pfam00535, Glycos_transf_2, glycosyltransferase, transferring sugar from UDP-glucose, UDP-N-acetyl- galactosamine, GDP-mannose or CDP-abequose, to a range of substrates including cellulose, dolichol phosphate and teichoic acids (score 76.8, E e^{-15}).

The second region selected for internal fragment replacement mutagenesis analysis was three ORFs near the cds29 insertion (Fig 1.20). Since these ORFs have been described earlier in this section, only a brief description of these mutagenized ORFs is given here. For more detailed information see the description for Cluster 1. The mutation 29-3 disrupts the orfC1-5-26, which does not have any database homologous sequence, but does have a G+C at the third base position of the codons at 90.5 %. Assuming translating from the first methionine, the peptide contains 328 amino acid residues. The mutation 29-5 disrupts the orfC1-5-24, which is homologous to the GumC (COG3206) domain (Score 47.8 E $3e^{-06}$), which is involved in exopolysaccharide biosynthesis. Its closest protein homolog is an uncharacterized protein (Score 63.9 E $1e^{-08}$) from *Nostoc punctiforme*, also carrying the GumC domain. The mutation 29-6 disrupts the glycosyltransferase domain-carrying orfC1-

5-23, which has strong homology (Score 162 E $2e^{-38}$) to a WcaA-like glycosyltransferase from *Desulfovibrio desulfuricans* G20.

We performed the following assays to characterize this set of six mutants: cohesion, development, and Calcofluor White-binding. In a standard cohesion assay, it usually takes 40 minutes or less for the wild type to complete the agglutination process, but for the polysaccharide deficient mutants it usually takes much longer or never goes to completion. Relative absorbance is calculated for tracing the progress of the cohesion assay from the ratio of the absorbance at a given time point to the initial absorbance of each cohesion assay reaction.

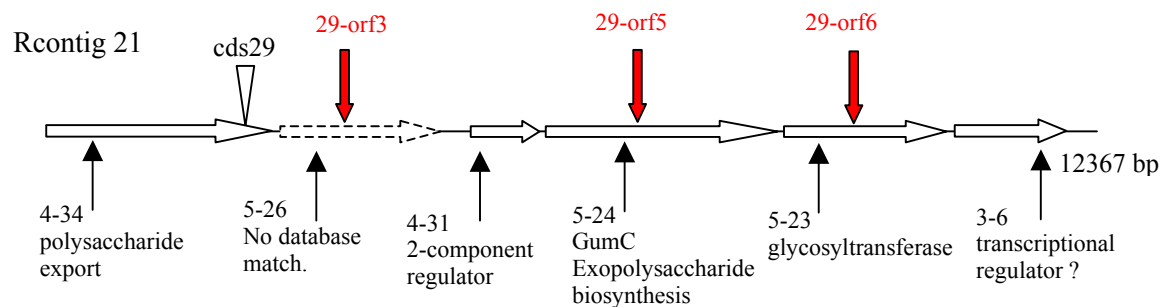


Figure 1.20. Sites of internal fragment replacement mutagenesis on near cds29 on Rcontig 21 at Cluster 1.

The data for strain 53-4 shows that it was able to complete the agglutination process, although it took much longer time (ca. 80 min. versus ca. 30 min.) compared to the wild type. The strains 53-3, 53-6, and all three new mutations near the cds29 are completely deficient in the cohesion assay. During the 80 minutes of incubation, the absorbance barely changed (figure 1.21). This result was not completely expected. What was unexpected was the orfC1-5-6, orfC1-5-4, and orfC1-6-2 are supposed to be in the same apparent operon. If

they were transcribed in a single transcript, interruption by 53-4 should have produced a polar effect on the gene downstream (interrupted by 53-3). Therefore, the result of 53-4 should have been more severely defective than that of 53-3, or at least both should have been about the same. However, the reverse is true. The interruption in the downstream ORF produced a more severe phenotype than in the upstream ORF. This phenomenon was seen in the truncation mutants in the *cds29* region as well. The last ORF in the apparent operon *orfC1-5-23* (interrupted by 29-6) produced the most severe defects in the development test. This means either that they are actually in two operons, or that the interruption introduced an artificial promoter into the upstream of the *orfC1-6-2*, and that negated the polar effects automatically. The first alternative interpretation has its own problem, however. If one pays close attention to the two ORFs' organization, the two ORFs have a four base pair overlap when considering the full length with high G+C bias as the coding sequence. The full length with high G+C starts from the first methionine, which overlaps the previous coding sequence (*orfC1-5-4*) by four bases. Assuming there is no promoter from the vector sequence, there must be either a native promoter inside the *orfC1-5-4* coding sequence, or a native promoter immediately downstream *orfC1-5-4*. This makes it possible that the initiation of translation may start from a CTG codon 20 codons downstream. The initiation codon cannot be further downstream because that would interrupt the conserved sequence found in the database sequences. Considering that the same phenomenon was observed in the interruption mutants near the *cds29* insertion, it is quite possible that there is a *M. xanthus* promoter in the pZerO-2 vector sequence. The possible promoters are the *Plac* promoter positioned to transcribe the cloning site from one direction, and the *Pkan^r* promoter positioned to transcribe the cloning site in the reverse direction. Among the two potential promoters, the *Pkan^r* promoter is known to be active in *M. xanthus* since the

mutants are kanamycin resistant. Assuming there is no terminator between the *kan^r* gene and the cloning site, the *Pkan^r* promoter is likely the source of promoter that breaks the polar affect. It is not known whether the *Plac* promoter is active in *M. xanthus*. This issue will be investigated further in future studies.

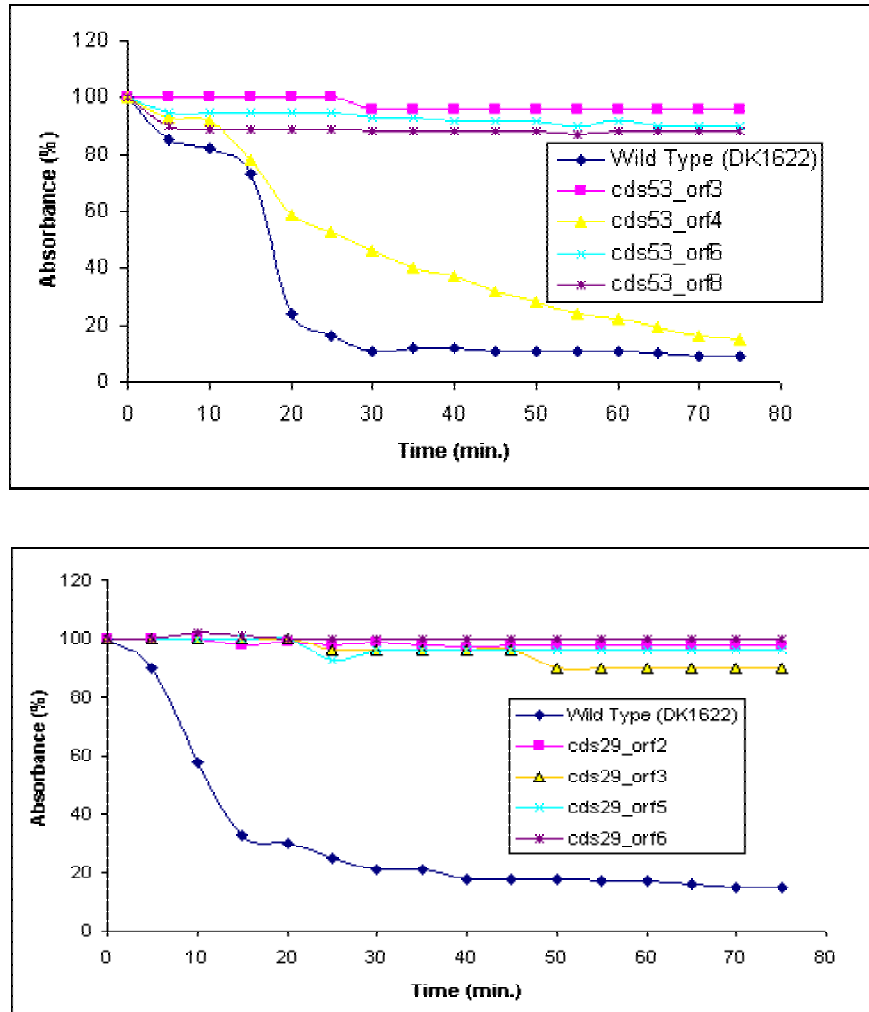


Figure 1.21 Cohesion comparison. The cohesion conditions are as specified in the Materials and Methods section. Note: The cds29-orf2 is in the same gene as cds29.

When assayed for their developmental characteristics, all the truncation mutants produced

developmental defects, albeit to different degree. Strain 29-2 is in the ORF that carries the original *cds29*. It has an interruption in the *orfC1-4-34*, which codes for a polysaccharide export system, possibly a GumB-like protein. The cells failed to aggregate into round mounds, and did not sporulate (Figure 1.22).

The developmental results show that *orfC1-5-26* (mutation 29-3) does encode a functional gene involved in the development process. Although mutant 29-3 forms aggregates, the aggregates were loose and failed to turn dark, indicating sporulation failure (Figure 1.22).

Strain 29-5 is interrupted in an exopolysaccharide biosynthesis gene GumC. Phenotypically its defects were similar to 29-2, probably reflecting the fact that they are two related proteins in the same exopolysaccharide export system. Strain 29-5 aggregated slightly poorer than 29-2. Aggregates from both strains stayed in the ridged form, and did not turn dark (Figure 1.22).

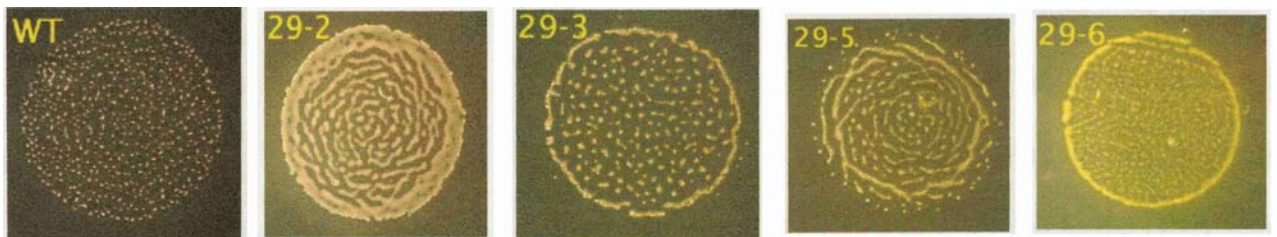


Figure 1.22 Development assay comparing strains carrying mutation in the *cds29* apparent operon. Highly concentrated cells were plated on TPM agar surface, and incubated at 30C for 72 hours. Wild type cells aggregated into defined fruiting bodies, and sporulated. Each mutant strain has a different degree of defects in the same developmental process.

Strain 29-6 endures an interruption in *orfC1-5-23*, which is a *wcaA*-like glycosyltransferase

gene. This interruption caused the most severe defect among all the targeted internal fragment replacement mutagenesis generated mutants. The defect suggested that the 29-6 cells fail to establish aggregation centers. There is very little aggregation, no sign for sporulation at all (Figure 1.22).

Strain 53-3 is interrupted in another glycosyltransferase. Compared with the strain 29-6, the glycosyltransferase interruption in 53-3 caused a mild defect. 53-3 is able to form aggregates, albeit abnormally broadly based and loosely defined. This probably reflects the fact that there is a transcriptional regulator downstream from the orfC1-5-23, while there is no ORF in sight downstream from 53-3 mutation. Strain 53-6 has an interruption in yet another glycosyltransferase. Similarly, the phenotype is very much the same as that of strain 53-3. The only difference is the strain 53-6 keeps better yellow coloration.

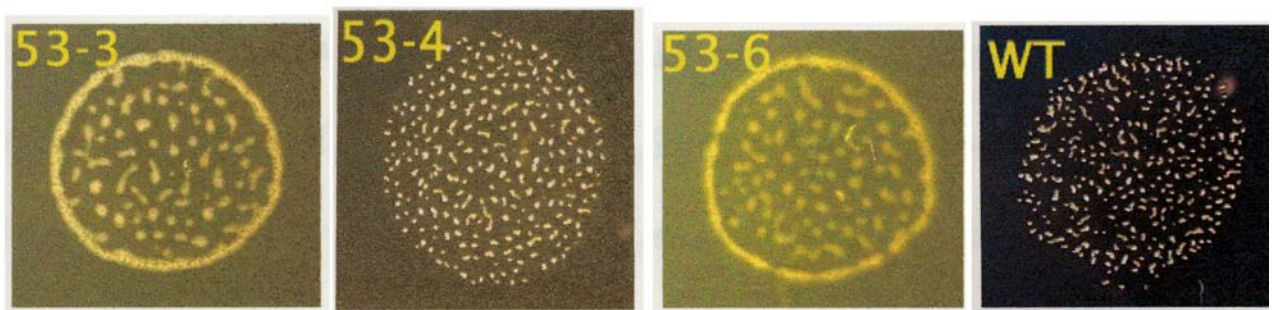


Figure 1.23 Development assay comparing strains carrying mutations near the *cds53* insertion. Highly concentrated cells were plated on TPM agar surface, and incubated at 30°C for 72 hours. Wild type cells aggregated into defined fruiting bodies, and sporulated. Each mutant strain has a different degree of defects in the same developmental process.

Strain 53-4 is interrupted in the *orf53-6-3*, which codes for an endoglucanase. This mutant displays the mildest defect among all internal fragment replacement mutagenesis generated

mutants. The agglutination assay shows this strain was able to complete the agglutination process, albeit it took a longer time. It formed aggregates almost normally under development conditions. However aggregates did not darken enough after 72 hours of development, indicating sporulation did not proceed normally. This observation coincides with A previous report about another endoglucanase beta-1,4-endoglucanase (CelA) from *M. xanthus* which is expressed during exponential growth, and also not involved in development either (Quillet *et al.*, 1995; Bensmail *et al.*, 1998).

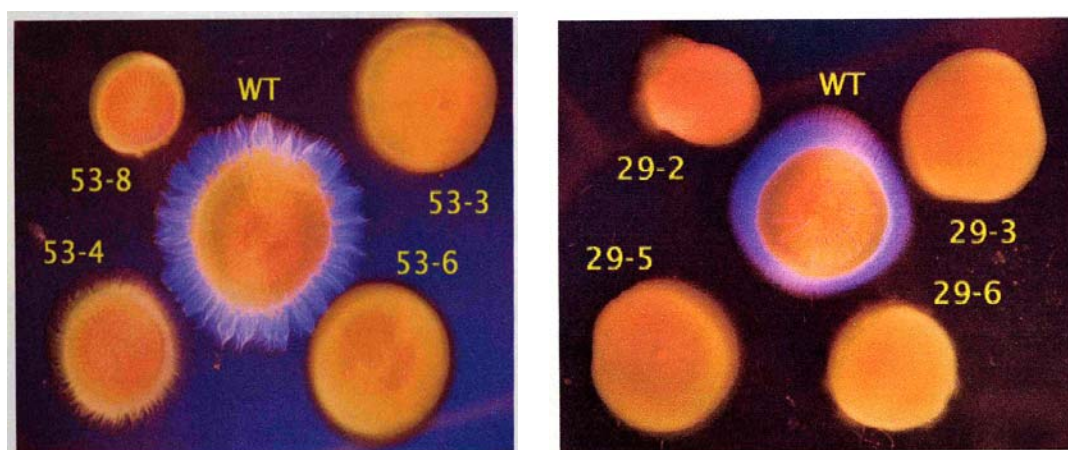


Figure 1.24 *M. xanthus* cells were plated on Calcofluor White containing plates. The wild type displays a highly fluorescent fringe while mutants show different degree of deficiency. The photo was taken with a combination of incandescent light and UV light (366 nm).

When plated on low percentage (0.3%) Calcofluor White-containing CYE agar plates (Fig. 1.24), the wildtype (DK1622) displayed a clear “hairy” S-motility fringe on the low percentage agar. The S-motility fringe was strongly fluorescent. All internal fragment replacement mutagenesis mutants were defective in Calcofluor White-binding. But it is obvious that the strain 53-4 did have a well-developed fringe, indicating its S-motility is not as severely impaired as in the other mutants. Interestingly, the S-motility fringe in the

mutant 53-4 showed no sign of fluorescence, suggesting that although the endoglucoanase interruption has little affect on its development, it is very important for Calcofluor White-binding and probably exopolysaccharide production. This seems to suggest that the correlation between the S-motility and development is stronger than the correlation between the S-motility and polysaccharide production. In other words, development capable strains more likely lose polysaccharide production capability than lose S-motility. This again verifies the earlier observation (Ramaswamy *et al.*, 1997) that strains that completely lost its ability to bind Calcofluor White still retain almost full amount of S-motility (Figures 1.5, 1.6, 1.7, 1.8 and Table 1.2).

CONCLUSION

Exopolysaccharide is a “catch all” category that includes lipopolysaccharide, peptidoglycan, and secreted polysaccharide. In short, any polysaccharide that is outside the cytoplasmic membrane is called exopolysaccharide. In *M. xanthus*, the basic units for polysaccharide include galactose, glucosamine, glucose, rhamnose, and xylose (Behmlander and Dworkin, 1994). The order of the basic units in a polysaccharide varies and is largely unknown. It is known that polysaccharide may contain a variety of branches. To synthesize polysaccharide often requires transferring the subunits through membrane-based enzymatic steps in addition to the final transportation to the outside of the cytoplasmic membrane. This makes it a very complex business to synthesize exopolysaccharides. A vast number of genes and families of genes have been identified, yet still an untold number of polysaccharide biogenesis genes have not been identified. Exopolysaccharide occurs in several forms in *M. xanthus*. One form is called fibrils on the cell surface (Kim *et al.*, 1999) which appear like a velvety coat. Another form looks like a spider’s “sticky strings” extending from the cell surface to any solid object in its vicinity (Behmlander and Dworkin, 1991). The third form is described as tactile sensors (Lee *et al.*, 1995). The fourth form is pictured as slime extruding from the polar nozzles functioning in A-motility (Wolgemuth *et al.*, 2002; Kaiser, 2003). All these forms are found to be important for motility and development. This dissertation project surveyed the genome of *M. xanthus* for potential genes involved in polysaccharide biosynthesis and development. Using colony morphology and Calcofluor White-binding properties, a collection of more than 40 mutants was selected from thousands generated. Among them, 25 were cloned. Of those cloned, 21 were sequenced and analyzed in this chapter. They are mapped to six clusters on the *M. xanthus* genomic sequence map.

The conclusions are drawn under the following eight topics.

Multiple Export Systems Involved In Exopolysaccharide Production

The most obvious feature is the involvement of type IV pilus system in the exopolysaccharide production. The type IV pilus system is known for its function in transportation of proteins. In fact, it shares many components with the type II protein secretion system (see Peabody *et al.*, 2003, for a recent review). Since the type II protein secretion system is believed to be a general secretory system, exopolysaccharide or the membrane-based enzymes, such as glycosyltransferases, could be exported through this system. The fact that every insertion (nine in total) in the apparent *pil* operon produces almost identical defects in Calcofluor White-binding and development strongly suggests that the type IV pilus system is indeed involved in exopolysaccharide biosynthesis and export. It is not clear at this point whether the exopolysaccharide synthesis and export related functions use the native type IV pilus system setup, or type II protein secretion system setup. It is also possible that the exopolysaccharide synthesis and export system uses yet another setup also sharing some components of the type IV pilus system.

In addition to the type IV pilus system components, a polysaccharide-specific export system with components such as GumB, GumC and RfbX may also be involved (Paulsen *et al.*, 1997; Hvorup *et al.*, 2003). The *cds29* insertion is in a *gumB*-homolog gene, and an *rfbX* gene is immediately downstream from the *cds32* and *cds33* insertions (Figure 1.11). These are known to be membrane proteins involved in polysaccharide export polysaccharide across both inner and outer membranes (Paulsen *et al.*, 1997). To summarize, although polysaccharide production is not well understood, at least we can conclude that

exopolysaccharide production in *M. xanthus* probably involves two export systems: a polysaccharide-specific export system and a “more general” export system. Since both export systems are required, they probably function in different ways. For example, type IV pilus system may function as an export system for glycosyltransferases that function only when properly positioned in a membrane and are required for polysaccharide biosynthesis or attachment; while the polysaccharide-specific export system (GumB, GumC and related proteins) may actually export the polysaccharide itself. The two export systems must operate in a well-coordinated way.

A Large Number Of Glycosyltransferases Exists In The *M. xanthus* Genome

There are six glycosyltransferases clustered in the Cluster 1. When searched with consensus domains of glycosyltransferases, tens more glycosyltransferase homologues were found in the *M. xanthus* genome. They were spread out over the whole genome. Glycosyltransferases are a class of enzymes not yet very well understood. Their functions, mechanisms of function, specificity, activation conditions, and classification are under intense investigation. For example, different sugar units may have different glycosyltransferases, such as glucosyltransferase and galactosyltransferase.

New Genes To Be Discovered

It seems that there are new polysaccharide production related genes to be discovered. The insertion cds1 identified an ORF in Cluster 3 with no homologue in the GenBank, yet with functions involved in polysaccharide production and development. Since the insertion is located at the end of an apparent operon, the coding sequence should be expressed and functional, suggesting that the cds1 insertion is in a real gene, although the gene does not

seem to be very much involved in motility.

Computer Assisted Sequence Analysis

Sequence analysis identified not only the locations where the insertions are located on the *M. xanthus* genomic contigs, but also allowed us to put the incomplete and unordered contigs into order and make a gapped genomic sequence map. On this genomic sequence map it is easy to show that there exist at least six clusters of genes involved in polysaccharide biogenesis. The most interesting feature is that the motility genes (*pil* and *dif* genes) are a major group of genes playing essential roles in exopolysaccharide biosynthesis and development.

The genomic sequence map assembled here is the first for the *M. xanthus* genome. The use of existing physical mapping data of the genome and the sizes of the restriction fragments provided the critical information to order and link the discontinuous contigs. A simple computer program to search the restriction sites and calculate the fragment sizes was a very useful tool for the job. This genomic sequence map not only presents the location and orientation of the contigs and the insertions (clusters of insertions), but also made it clear where (between which contigs) the gaps are and roughly how big the gaps are. Therefore this approach of assembling a genomic sequence map has good potential to be used to close final sequencing gaps in genomic sequencing projects.

For mapping the insertion sites, another set of computer programs called SHAPE was developed to automate the process. Although the number of our insertion sequences is limited, since the *M. xanthus* genome sequence has been an on-going project, repeatedly

manually mapping the insertions would be a waste of efforts. Besides, this set of programs can be used for any sequence mapping projects and is going to be available to the research community. The potential benefit is worth the effort. A demonstration web site is at <https://129.15.160.110/DNAmapping/> (Note: the last slash is required). Since the set of programs are designed to present graphical, global alignment, rather than a typical BLAST-kind of alignment, it provides a lot more convenience to the user. See Chapter 2 for details.

Polysaccharide Production Is Correlated With Smooth Colony Morphology

Many genes found in this dissertation were studied before; however, the characteristics of polysaccharide production deficiencies are rarely discussed. For example, before this project was undertaken, it was essentially unknown that the *pil* mutants are all deficient in exopolysaccharide production. The availability of an incomplete genomic sequence database helped to piece together the large clusters, in the range of one hundred kb or more. As a result, the complete apparent *pil* operon has been revealed for the first time. This in turn helped to explain some of the phenotypes and their relationships, particularly in the *pil* locus. (Note: The riboflavin synthase in front of *pil* genes seems to be very interesting. It might indicate that FAD is involved in pilus function or production.) We found that the *tgl* gene is in fact homologous to the domain of the *pilF* gene, which is required for pilus assembly and coincides with known functions for the *tgl* gene. However, our understanding about *pilF* is quite limited. The fact that the PilF protein in *M. xanthus* is an outer membrane protein (Simunovic *et al.*, 2003) and can be supplied externally from neighboring cells (Hodgkin and Kaiser, 1979; Wall *et al.*, 1998) implies that it can translocate from one membrane to another without losing its functionality. This property is probably related to conjugation of the type IV pilus system, and could be investigated for delivering

therapeutical via conjugation vectors.

This survey made it possible to correlate smooth-looking colony morphologies with a deficiency in exopolysaccharide biosynthesis, and development. All smooth-looking strains used in this study showed defects in polysaccharide production and development. It is also clear that there are many well-studied developmental genes (*pil* genes, for example) that in fact are required for exopolysaccharide synthesis, transportation, and/or assembly. In addition, our evidence shows that all 9 insertion mutants in the *pil* cluster are defective not only in social motility, but also in polysaccharide production. Therefore, we propose that the pil-like proteins, especially the secretin PilQ, are involved in polysaccharide export. Since the characteristics of polysaccharide production have not been routinely examined in developmental mutants, we believe that the effect of polysaccharide production on developmental defects could be more extensive than we can estimate at this time.

***M. xanthus* Tends To Intersperse Environment-Sensing Genes Into Polysaccharide Production Genes To Presumably Coordinate Their Activity**

A BLAST search using conserved domains for two-component signal transduction systems found that *M. xanthus* has more than 50 homologs with E value less than e^{-30} . The two-component signal transduction elements occur in all six insertion clusters discussed in this work, and at high density at times. For example, in the *pil* locus (Cluster 5) there are two pairs plus a CheY-like element. We believe this could be a mechanism for *M. xanthus* to couple polysaccharide production and other activities with developmental signals. Without the information about the promoter and other regulatory components, it is difficult to

understand the organization. But the result from strain 29-6 seems to support this hypothesis. As discussed above, strain 29-6 has an interruption in the orfC1-5-23, similar to strain 53-3. However, 29-6 shows a much more severe developmental defect. The explanation could lie in the differences between the two glycosyltransferase genes. However, it is equally possible that the surrounding genes made the difference. One obvious possibility in this aspect is the polar effect of the insertional interruption of the glycosyltransferase. The severity could be due to the elimination of the downstream transcriptional regulator. Another piece of supporting evidence came from an earlier study on the SR53 insertion (Ramaswamy *et al.*, 1997). They showed severe defects in SR53's development, Calcofluor White-binding, and agglutination. We know now that downstream from the glycosyltransferase that the SR53 insertion disrupted is a two-component hybrid sensor gene. Using the same reasoning as above, the defect probably is due to the polar effects of the insertion SR53, which prevented the expression of the 2-component sensor.

Future Mutagenesis Analysis

Among the six exopolysaccharide production related clusters, the *pil* gene locus received an unusually high percentage of insertions. This is probably due to several reasons. One, the *pil* locus is a large locus, more than 23 kb in length. A large locus naturally receives a large number of random insertion events just by chance. Two, the insertion mutants in this locus all display a dramatic Calcofluor White-binding deficiency. Therefore, insertions in this area are more likely to be selected for analysis. We know that the Calcofluor White-binding deficiency is correlated to polysaccharide production. This is the intentional bias. Only after a substantial number of the same random transposon insertion experiments can we say whether the bias we see is a reliable reflection of the reality in the genome.

Although the *pil* genes are the most frequently hit by the *magellan-4* transposon, the frequency of more than one hit per gene was very low. Many genes in the locus were not hit. This means that there are other polysaccharide production genes still not identified. A preliminary conserved domain search in the incomplete *M. xanthus* genome found tens of glycosyltransferase homologs scattered in the genome. This exactly coincides with the complex exopolysaccharide biosynthesis needs as explained above. Consequently it is most likely that there are more genes to be found for polysaccharide production and development by this same technique. Employing other techniques to find mutants could increase the search potential.

Since FB is a product of three base changes, we expect more genes will produce smooth phenotypes when using single base change mutagenesis, such as using UV irradiation, chemical and spontaneous mutagenesis to generate mutants. This class of methods will probably discover some genes that cannot be discovered using the transposon methods, such as the genes which encode or make an integral membrane element (proteins, peptidoglycan, etc.) that may be essential for the cell's viability. However, screening and cloning these genes can be more difficult.

Other observations

All the internal fragment replacement mutagenesis mutants are generated using the same plasmid vector, therefore probably carry similar side effects. The design of internal fragment replacement mutation is to truncate the target gene, and the genes downstream from the target site are expected to suffer a polar effect. However, two sites (near SR53 and

cds29) of truncation mutagenesis did not show much of polar effect on the genes downstream in the apparent operons. This could mean that the plasmid vector used (pZerO-2) probably carries some *M. xanthus* recognizable promoter, such as *Plac* or *Pkan^r* promoters. If this is proven true, the pZerO-2 vector could be quite useful in some studies, but for our initial design, a different vector should be used.

Finally, the acronym *cds* stands for Calcofluor White-binding deficient and S-motile (Ramaswamy *et al.*, 1997). However, the motility data collected here show that this group of *cds* mutants has a whole range of S-motility deficiencies. Consequently, while Calcofluor White-binding deficiency is the key common character for the group, the S-motility varies greatly. Therefore, the name itself has accumulated some extra meaning over time. It is proposed to give the acronym *cds* a new definition: Calcofluor White-binding deficient and S-motility variable.

CHAPTER 2

A COMPUTER PROGRAM FOR AUTOMATED INSERTION POINT MAPPING

INTRODUCTION

There are many computer programs for biological sequence analysis. Usually each program facilitates one aspect of sequence analysis. Therefore, researchers have to use many sequence analysis programs to complete an analysis, even when a comprehensive program package is used. In other words, sequence analysis encompasses a wide range of subjects. Sequence analysis could mean DNA sequence analysis, RNA sequence analysis, or amino acid sequence analysis. There are many different analyses even in DNA sequence analysis alone. For example, one may be interested in analyzing the DNA for open reading frames, conserved domains, primers for PCR, repeated sequence fragments, secondary structures, promoters, terminators, and so on. A vast number of programs have been developed for these tasks. Don Gilbert (Gilbert, 2000) listed 160 of them freely available in the year 2000. Many of them are web accessible to all users, such as BLAST, CLUSTAL, and Artemis. There are an untold number of others not web accessible, or with restricted web access, such as GCG.

Since my goal is to develop a computer program to help me find the insertion point of transposon insertions in the *M. xanthus* genome. The focus for this chapter is on DNA sequence analysis tools. In this chapter, only DNA-related computer programs, or DNA-

related functions of a program tool are discussed, other tools and functions are ignored. One variety of sequence analysis program is the sequence alignment program. Probably the best known of them is BLAST (Altschul et al., 1990; Cummings et al., 2002; Mthog 2003; Korf 2003). BLAST searches databases for homologous sequences to a given query sequence and generates alignments once significant homologous sequences are found in the databases. An example of the alignments follows:

```

Query: 14   tacgcccaccccaggtttccctgggcgatgaaggcgctcgggtagagcatgtcattgagg 73
          |||
Sbjct: 1343 tacgcccaccccaggtttccctgggcgatgaaggcgctcgggtagagcatgtcattgagg 1284

Query: 68   cgtcggcgctatgttgacacgctccgccaactccttcttcttctgtacgtcggaggcgct 127
          |||
Sbjct: 3299 cgtcggcgct-tgttcgcgaactccgccagctccttcttcttctgt-cgtcgggaagcgct 3242

Query: 128  cgactggcgctccgaaatcatcgccctggatttccgacctatacaggcacggagttgctc 187
          |||
Sbjct: 3241 cgactggcgctccgaaatcatcgccctggatttccg-cctcggacaggc-cggagttgctc 3184

Query: 188  accacgcgcaccagctgaaccttgccggcgccctacgtccttggcgctgaactggacgat 247
          |||
Sbjct: 3183 accacgcgcacctgctgaaccttgccgggtgcc-acggccttggcgctgacgtggacgat 3125

Query: 248  gcccatggcgctccatgtcgaacgacacctcgaattgcggcacgacctcggcgccctcgc 307
          |||
Sbjct: 3124 gccgttggcgctcgatgtcgaacgacacctcgatttgcggcacg-cgcgcggcgccgggg 3066

Query: 308  gaatgccaccatttcgaagcgcgccacctcttgttgcgcgccatctcacgctcgc 367
          |||
Sbjct: 3065 gaatgccaccagttcgaagcgcgccacctcttgttgcgcgccatctcacgctcgc 3006

Query: 368  cctggagcacgtgcacgctcaccagcggtggttgcacggcggtggagaacacctgnc 427
          |||
Sbjct: 3005 cctggagcacgtgcacgctcaccagcggtggttgcacggcggtggagaacacctggc 2946

Query: 428  tcttcttgacgaaataatggcgcttcttgcgaatcatttctgtaaacacaccgcc 482
          |||
Sbjct: 2945 tcttcttgacgggatggtggtgttcttgcgatgatttctgtaaacacaccgcc 2891

```

Figure 2.1. Alignment of two DNA sequences. Output from BLASTN, showing the cds40_mar1 against the GenBank sequence at the *M. xanthus* sglK Cluster.

According to the scope of the alignment, sequence alignments can be categorized into global and local alignments. Local alignment uses fragments of the query sequence search against the subject sequence (database sequence). Once a significant match is found BLAST tries to extend the homology from both ends of the matching query fragment. The BLAST output is a local alignment. One prominent feature of a BLAST output is that the flanking

regions (non-homologous parts) are not included. If there is more than one stretch of alignment in the neighborhood, the neighboring alignments will be presented in separate blocks. BLAST output breaks the integrity of the sequence into pieces (lines), therefore it may limit the detection of continuous sequence similarity (Fig. 2.1). Local alignment tries to address the question of what the query sequence codes for based on similarity to known gene sequence in the databases.

There are many specialized versions of BLAST in use. The BLASTN program uses DNA query sequences to search the DNA databases, while the BLASTP program uses protein sequences to search the protein databases. Since biologists are more interested in what the DNA codes for and the functions of the coded proteins (using BLASTP and/or BLASTX), rather than to find out whether a piece of novel DNA sequence is discovered, BLASTN is not as widely used by biologists. Consequently, there isn't very much research or development activity to extend the nucleotides-nucleotides alignment functions. The most notable use of nucleotides-nucleotides alignment is in sequencing facilities. Nowadays, sequencing is highly automated, including the nucleotides-nucleotides alignment and assembly processes, which becomes an integral part of the automated sequencing. There is no human intervention necessary to align and assemble the sequences. The only human interaction is involved in the quality checking the assembled sequences. Therefore those programs in the sequencing machines are not user accessible. Even if the programs for matching and assembling are human accessible, they are proprietary in nature, their use will be restricted. In sharp contrast, protein-protein alignment is still at a stage that human interpretation is indispensable. At the present, an intense effort is focused on the protein sequence analysis (Higgins et al., 1996, Yuan et al., 1999, Campagne 200, Chenna et al.,

2003, Gasteiger et al., 2003).

Global alignment is used when one wants to know where a piece of ones' favorite DNA is in the context of a long continuous sequence (contig) or relative to another piece of ones' favorite DNA. A global alignment addresses the question of overall similarity and /or distribution (position) of one sequence relative to the other. Global alignment uses the whole query sequence to search against and align with the whole subject sequence. Global amino acid sequence alignment is limited to the length of the peptide, which is usually less than 1000 amino acids long. But global alignment of DNA sequences deals with size ranges from a few bases to many millions of bases, the size of a chromosome. Global DNA alignment is needed if one has several insertion mutants in a cluster, and wants to know where exactly in that cluster each of the insertions is. This knowledge probably will enable one to correlate the phenotypes to a specific gene or genes.

BLAST search results are often broken into separate alignment blocks. It is much more frequent in protein sequence alignment, but still occurs in DNA sequence alignment. The following example is a small contig (contig432) in the *Myxococcus xanthus* genome database search with BLASTN against the whole genome of *M. xanthus*:

```
>gnl|TIGR_246197|contig:526:m_xanthus Myxococcus xanthus DK 1622 unfinished fragment of genome
Length = 3581774

Score = 5909 bits (3073), Expect = 0.0
Identities = 3073/3073 (100%)
Strand = Plus / Plus

Query: 1      gttccgggcatggtgagcaccggttccgggcatggtgagcaccgattccggcccgaggtg 60
            |||
Sbjct: 3098827 gttccgggcatggtgagcaccggttccgggcatggtgagcaccgattccggcccgaggtg 3098886

Query: 61      agcaccggttccgggcatggtgagcaccggttccggacatggtgagcacagtcgggaacg 120
            |||
Sbjct: 3098887 agcaccggttccgggcatggtgagcaccggttccggacatggtgagcacagtcgggaacg 3098946

Query: 121     gtacggagcggttgacggcaactgggcagcaggggtctcgctccacctccttcgaggaggtg 180
            |||
Sbjct: 3098947 gtacggagcggttgacggcaactgggcagcaggggtctcgctccacctccttcgaggaggtg 3099006

Query: 181     gagatggccaagagaggtggcggtgcgcaagttgagagaggtgttcggttcggttc 240
            |||
Sbjct: 3099007 gagatggccaagagaggtggcggtgcgcaagttgagagaggtgttcggttcggttc 3099066
```

Query: 241 gcgtcgaagctgtcgacgaggaacatcgccacgagctctgggcatagggaatgggacggtg 300
 |||||
 Sbjct: 3099067 gcgtcgaagctgtcgacgaggaacatcgccacgagctctgggcatagggaatgggacggtg 3099126

Query: 301 tgcgagtagcctggggcgagcgcgggtagcagggtgggagactggccgctgccgcggag 360
 |||||
 Sbjct: 3099127 tgcgagtagcctggggcgagcgcgggtagcagggtgggagactggccgctgccgcggag 3099186

Query: 361 ctggacgacgacgcggcgctcaccgcgctctcttccctgccgagggcaagggggttgcg 420
 |||||
 Sbjct: 3099187 ctggacgacgacgcggcgctcaccgcgctctcttccctgccgagggcaagggggttgcg 3099246

Query: 421 caccggccggagccggactggcgcgagtgcatcgagagctcaagcgaaaggggttacc 480
 |||||
 Sbjct: 3099247 caccggccggagccggactggcgcgagtgcatcgagagctcaagcgaaaggggttacc 3099306

Query: 481 aagctgctgttggggaggtagtacctggcgccaaccgggtgggtaccagtacagccag 540
 |||||
 Sbjct: 3099307 aagctgctgttggggaggtagtacctggcgccaaccgggtgggtaccagtacagccag 3099366

Query: 541 ttttgcgagcgggtatggcgctggcagtcctgctcgtgtccaccatgagacaggagcac 600
 |||||
 Sbjct: 3099367 ttttgcgagcgggtatggcgctggcagtcctgctcgtgtccaccatgagacaggagcac 3099426

Query: 601 cgcgcggggcgagaagctctctgtggaacttcagcggggatggagtcgaggtggtggagcgc 660
 |||||
 Sbjct: 3099427 cgcgcggggcgagaagctctctgtggaacttcagcggggatggagtcgaggtggtggagcgc 3099486

Query: 661 gacaccggagaagtgcgggtagcgaagctcttcgtgccacgctggggccagcagttac 720
 |||||
 Sbjct: 3099487 gacaccggagaagtgcgggtagcgaagctcttcgtgccacgctggggccagcagttac 3099546

Query: 721 acgtacgtcgagcccgcttactccgaggatttgccacactgggtgggtgccacgtgcgc 780
 |||||
 Sbjct: 3099547 acgtacgtcgagcccgcttactccgaggatttgccacactgggtgggtgccacgtgcgc 3099606

Query: 781 gccatggcctctcttggcggtactccggcggttggtggtgccgacaaactgaagtcgggc 840
 |||||
 Sbjct: 3099607 gccatggcctctcttggcggtactccggcggttggtggtgccgacaaactgaagtcgggc 3099666

Query: 841 gtcacccacgtgcaccgctacgagccggaggagaatcccacgtacgccacactggcccgg 900
 |||||
 Sbjct: 3099667 gtcacccacgtgcaccgctacgagccggaggagaatcccacgtacgccacactggcccgg 3099726

Query: 901 cactacggcttcgccattctgcggcgcgctcctgcgcgccgcgcgacaaagcggaagtg 960
 |||||
 Sbjct: 3099727 cactacggcttcgccattctgcggcgcgctcctgcgcgccgcgcgacaaagcggaagtg 3099786

Query: 961 gaggccgcggtgctggtggctcagcggtggattctgcccgtcctgcgaaccaccgcttc 1020
 |||||
 Sbjct: 3099787 gaggccgcggtgctggtggctcagcggtggattctgcccgtcctgcgaaccaccgcttc 3099846

Query: 1021 ggtggcctgcacgaggtacgtgagggcgtacggccgttgctcgagaagctgaatggccgc 1080
 |||||
 Sbjct: 3099847 ggtggcctgcacgaggtacgtgagggcgtacggccgttgctcgagaagctgaatggccgc 3099906

Query: 1081 ccgatgcggcatgtggggcgctcgcgtcgccagctgtacgaggagctcgagaagcctgtg 1140
 |||||
 Sbjct: 3099907 ccgatgcggcatgtggggcgctcgcgtcgccagctgtacgaggagctcgagaagcctgtg 3099966

Query: 1141 ctgaaggccctgcgggtacacgcctacgagctggcctctctggaagaaggcgcgctccac 1200
 |||||
 Sbjct: 3099967 ctgaaggccctgcgggtacacgcctacgagctggcctctctggaagaaggcgcgctccac 3100026

Query: 1201 cctgactaccacgtcgaggtggaggggcacctctacagcgtgccgtactcgtggcgcac 1260
 |||||
 Sbjct: 3100027 cctgactaccacgtcgaggtggaggggcacctctacagcgtgccgtactcgtggcgcac 3100086

Query: 1261 aagcaggtggagcccgcgtacacggaggggagcgtcgaggtgttccctcgggggccgctcg 1320
 |||||
 Sbjct: 3100087 aagcaggtggagcccgcgtacacggaggggagcgtcgaggtgttccctcgggggccgctcg 3100146

Query: 1321 gtcgccagccacgtgcgcaagcacgccaagggtacaccacgctgaaggagcacatgcc 1380
 |||||
 Sbjct: 3100147 gtcgccagccacgtgcgcaagcacgccaagggtacaccacgctgaaggagcacatgcc 3100206

Query: 1381 gccagccacggggcccacgcggagtgaagccacgcggctgctgacatggggcgagaag 1440
 |||||
 Sbjct: 3100207 gccagccacggggcccacgcggagtgaagccacgcggctgctgacatggggcgagaag 3100266

Query: 1441 acgggcccttcacggcgcgcttggtgcaaggcctcatggagcgaaaaccccatccggag 1500
 |||||
 Sbjct: 3100267 acgggcccttcacggcgcgcttggtgcaaggcctcatggagcgaaaaccccatccggag 3100326

Query: 1501 cagggtctccgcggggccttggtgtcatcgattgaaggacaagtacggagagcgcg 1560
 |||||
 Sbjct: 3100327 cagggtctccgcggggccttggtgtcatcgattgaaggacaagtacggagagcgcg 3100386

Query: 1561 ctggagaaggcgtgcgccagggcagtgctcacggggcctacagctacaagtcogtggcc 1620
 |||||
 Sbjct: 3100387 ctggagaaggcgtgcgccagggcagtgctcacggggcctacagctacaagtcogtggcc 3100446

Query: 1621 gccatcctccagcaccacctggaggacgcgcgggaggagcgcgaggagaagccgccctg 1680
 |||||
 Sbjct: 3100447 gccatcctccagcaccacctggaggacgcgcgggaggagcgcgaggagaagccgccctg 3100506

Query: 1681 ccgcgccatgagaatgtgcgggccccactactaccactgacgtacctctcgcgtccg 1740
 |||||
 Sbjct: 3100507 ccgcgccatgagaatgtgcgggccccactactaccactgacgtacctctcgcgtccg 3100566

Query: 1741 cgaggtgcaccgcgccacactgcgcgggtggagcccttggtgcgcggaacattcccg 1800
 |||||
 Sbjct: 3100567 cgaggtgcaccgcgccacactgcgcgggtggagcccttggtgcgcggaacattcccg 3100626

Query: 1801 ctggtgcgagtcctcggtcacccgcacccggaagggggcccgcgcgccctgagaagctg 1860
 |||||
 Sbjct: 3100627 ctggtgcgagtcctcggtcacccgcacccggaagggggcccgcgcgccctgagaagctg 3100686

Query: 1861 gcgcgcgcgggcaagcccccatgaccgacatgaggaacgaagccaatgctggtggaacag 1920
 |||||
 Sbjct: 3100687 gcgcgcgcgggcaagcccccatgaccgacatgaggaacgaagccaatgctggtggaacag 3100746

Query: 1921 acgctggagaaactcaacgggatgaagctgcacgggatggcctcgtaacctgcgcgactgg 1980
 |||||
 Sbjct: 3100747 acgctggagaaactcaacgggatgaagctgcacgggatggcctcgtaacctgcgcgactgg 3100806

Query: 1981 ttggcgaggccaggggagcgagatgttgccccagcgacactggtgggctgctggccgac 2040
 |||||
 Sbjct: 3100807 ttggcgaggccaggggagcgagatgttgccccagcgacactggtgggctgctggccgac 3100866

Query: 2041 gcggagtggatgcaccgagagaacaagaactctcctctcggtcgagccgcgcgcctg 2100
 |||||
 Sbjct: 3100867 gcggagtggatgcaccgagagaacaagaactctcctctcggtcgagccgcgcgcctg 3100926

Query: 2101 cgccaggccgcgcgccttggaagacatcgactacgggcacgcgcgcgggctcgcaaaagact 2160
 |||||
 Sbjct: 3100927 cgccaggccgcgcgccttggaagacatcgactacgggcacgcgcgcgggctcgcaaaagact 3100986

Query: 2161 caggtgatggagctgtccacctcgaagtggcgcgacaaagcagaatgctcctcctacc 2220
 |||||
 Sbjct: 3100987 caggtgatggagctgtccacctcgaagtggcgcgacaaagcagaatgctcctcctacc 3101046

Query: 2221 gggcccccaggcgctgggcaaatccttctcgcacgcgcctgggcccagaaggcgtgtcg 2280
 |||||
 Sbjct: 3101047 gggcccccaggcgctgggcaaatccttctcgcacgcgcctgggcccagaaggcgtgtcg 3101106

Query: 2281 gatggctactcggtggtgtaccgcgggacctcactctcttgcgatgagctcgccaggcg 2340
 |||||
 Sbjct: 3101107 gatggctactcggtggtgtaccgcgggacctcactctcttgcgatgagctcgccaggcg 3101166

Query: 2341 cgcgccgatggaacctacgcgcacgtgctcaagcgactggccaaggccaggtgctcacc 2400
 |||||
 Sbjct: 3101167 cgcgccgatggaacctacgcgcacgtgctcaagcgactggccaaggccaggtgctcacc 3101226

Query: 2401 ctcgatgacttcggccttgagccgctcggcgctccggagcgcaaggagtgtctcgaagt 2460
 |||||
 Sbjct: 3101227 ctcgatgacttcggccttgagccgctcggcgctccggagcgcaaggagtgtctcgaagt 3101286

Query: 2461 ctggaggaccgctaccagctcgcgagcacctgggtgacatcccgacttgagccgaaagac 2520
 |||||
 Sbjct: 3101287 ctggaggaccgctaccagctcgcgagcacctgggtgacatcccgacttgagccgaaagac 3101346

Query: 2521 tggcacgcgctcatcggcgcacgcgcgctcgccgcgcacatcctcgacgctctggtccac 2580
 |||||
 Sbjct: 3101347 tggcacgcgctcatcggcgcacgcgcgctcgccgcgcacatcctcgacgctctggtccac 3101406

Query: 2581 aacgcccatcgcatcaagctgggcggagagtccatccggtacgtggagacaaatttgacg 2640
 |||||
 Sbjct: 3101407 aacgcccatcgcatcaagctgggcggagagtccatccggtacgtggagacaaatttgacg 3101466

Query: 2641 aaggggccgcaagcaggccaagggatgaaccaccacgcgtcgctgacgctccgaccgctcg 2700
 |||||
 Sbjct: 3101467 aaggggccgcaagcaggccaagggatgaaccaccacgcgtcgctgacgctccgaccgctcg 3101526

Query: 2701 ccatcagccggaatcgctgctcggttgagccggaacgagtgtccacctgaccggaata 2760
 |||||
 Sbjct: 3101527 ccatcagccggaatcgctgctcggttgagccggaacgagtgtccacctgaccggaata 3101586

Query: 2761 cgcactcccacctcctcaggctgggtagcgtcgccagcgccggccattcggtgcgcg 2820
 |||||
 Sbjct: 3101587 cgcactcccacctcctcaggctgggtagcgtcgccagcgccggccattcggtgcgcg 3101646

Query: 2821 cagggtggtgtgacggaggaatggattggcgctccgcagacgtgcaagccggaatgga 2880
 |||||
 Sbjct: 3101647 cagggtggtgtgacggaggaatggattggcgctccgcagacgtgcaagccggaatgga 3101706

Query: 2881 cggtggtgcatacgcaggtggcctcgctcgctcggtcccgcgacacctgcgtggtggc 2940
 |||||
 Sbjct: 3101707 cggtggtgcatacgcaggtggcctcgctcgctcggtcccgcgacacctgcgtggtggc 3101766

Query: 2941 ctgggcccgcgcggcgccacgtcacctccaaggccctggcgggggagactgatgcgt 3000
 |||||
 Sbjct: 3101767 ctgggcccgcgcggcgccacgtcacctccaaggccctggcgggggagactgatgcgt 3101826

Query: 3001 cactttccgcgcgcgtccccggtaacgcggcctggcgctcgccgtgcgcacgcgtccgt 3060
 |||||
 Sbjct: 3101827 cactttccgcgcgcgtccccggtaacgcggcctggcgctcgccgtgcgcacgcgtccgt 3101886

Query: 3061 gcgggtgcgctgg 3073
 |||||
 Sbjct: 3101887 gcgggtgcgctgg 3101899

Score = 1904 bits (990), Expect = 0.0
 Identities = 990/990 (100%)
 Strand = Plus / Plus

Query: 3135 agtagcgctgccccctgtcgacacggctgcacagctggaacagcctccggcaactccggcc 3194
 |||||
 Sbjct: 3101961 agtagcgctgccccctgtcgacacggctgcacagctggaacagcctccggcaactccggcc 3102020

Query: 3195 gcccgacaccaccagacgcaatgactccacgcacacctcccgtgcgcgcgacttctg 3254
 |||||
 Sbjct: 3102021 gcccgacaccaccagacgcaatgactccacgcacacctcccgtgcgcgcgacttctg 3102080

Query: 3255 gcgccgcgcgcggagtgcgctaccccttctcaccggctcgccctctcgacacaccagg 3314
 |||||
 Sbjct: 3102081 gcgccgcgcgcggagtgcgctaccccttctcaccggctcgccctctcgacacaccagg 3102140

Query: 3315 gcgcgccccaggaggccctccaggtgcgtggtggcctcttctgacttttcaggagcg 3374
 |||||
 Sbjct: 3102141 gcgcgccccaggaggccctccaggtgcgtggtggcctcttctgacttttcaggagcg 3102200

Query: 3375 gcgcagcccccagactccacggctcgcgctcctctcctcaacttttcggcgccaccagc 3434
 |||||
 Sbjct: 3102201 gcgcagcccccagactccacggctcgcgctcctctcctcaacttttcggcgccaccagc 3102260

```

Query: 3435      gctgcccctggcagcgagctctcgaacttctcgcgctccatgaccagaaatccgcgctt 3494
                |||
Sbjct: 3102261  gctgcccctggcagcgagctctcgaacttctcgcgctccatgaccagaaatccgcgctt 3102320

Query: 3495      cctggtgaccatcaacacccgatgaaggcgctacgctcttgccaccttgacgtggctg 3554
                |||
Sbjct: 3102321  cctggtgaccatcaacacccgatgaaggcgctacgctcttgccaccttgacgtggctg 3102380

Query: 3555      gggtcaacccatcttcaagcgcccgcttccggacgacaatgcgttctccgaggcc 3614
                |||
Sbjct: 3102381  gggtcaacccatcttcaagcgcccgcttccggacgacaatgcgttctccgaggcc 3102440

Query: 3615      ctctccgcacgctgaaataccgccccaccttccgcagcgcccttcgcgtccgtcgag 3674
                |||
Sbjct: 3102441  ctctccgcacgctgaaataccgccccaccttccgcagcgcccttcgcgtccgtcgag 3102500

Query: 3675      gacgcgctgcatgggtgatgcgcttcatggcttggtacaacagcgagcacggcactcc 3734
                |||
Sbjct: 3102501  gacgcgctgcatgggtgatgcgcttcatggcttggtacaacagcgagcacggcactcc 3102560

Query: 3735      gccatccgcttcgtcaagcgacgacagacattccggccgagggcaacgctcctcgcc 3794
                |||
Sbjct: 3102561  gccatccgcttcgtcaagcgacgacagacattccggccgagggcaacgctcctcgcc 3102620

Query: 3795      cggcgcgaccaagtgtatctgcgtgccgagtcgctcaccgcgagcgctggagaggtggc 3854
                |||
Sbjct: 3102621  cggcgcgaccaagtgtatctgcgtgccgagtcgctcaccgcgagcgctggagaggtggc 3102680

Query: 3855      acacgcaactggacgccagcgcccgctcgtctccggccctctccgaacctctccccc 3914
                |||
Sbjct: 3102681  acacgcaactggacgccagcgcccgctcgtctccggccctctccgaacctctccccc 3102740

Query: 3915      gcagaacaggagatgaagcgcatcggtgaacccctcaccggtcggtcgggggcgcg 3974
                |||
Sbjct: 3102741  gcagaacaggagatgaagcgcatcggtgaacccctcaccggtcggtcgggggcgcg 3102800

Query: 3975      cgcggtggtgtgcaggctgctcttggcgatatccgcgtttccggcttcgtgcccggtga 4034
                |||
Sbjct: 3102801  cgcggtggtgtgcaggctgctcttggcgatatccgcgtttccggcttcgtgcccggtga 3102860

Query: 4035      gagcgcccatcactccgtctccggtcccgccacatcgaacgggacaggcggtttccc 4094
                |||
Sbjct: 3102861  gagcgcccatcactccgtctccggtcccgccacatcgaacgggacaggcggtttccc 3102920

Query: 4095      gcatccggtcaccgcgaaggcgctcatctc 4124
                |||
Sbjct: 3102921  gcatccggtcaccgcgaaggcgctcatctc 3102950

Score = 327 bits (170), Expect = 9e-89
Identities = 170/170 (100%)
Strand = Plus / Minus

Query: 3955      gcggtcgggtcgggggcccgcggtggtgtgcaggctgctcttggcgatatccgcgtt 4014
                |||
Sbjct: 804889  gcggtcgggtcgggggcccgcggtggtgtgcaggctgctcttggcgatatccgcgtt 804830

Query: 4015      tccggttcgtgcccggttagagcgcccatcactccgtctccggccccccacatcga 4074
                |||
Sbjct: 804829  tccggttcgtgcccggttagagcgcccatcactccgtctccggccccccacatcga 804770

Query: 4075      accggacaggcggtttcccgcatccggctcaccgcgaaggcgctcatctc 4124
                |||
Sbjct: 804769  accggacaggcggtttcccgcatccggctcaccgcgaaggcgctcatctc 804720

```

Figure 2.2. BLAST matches and aligns sequences in segments, breaks human intuition about the query sequence's wholeness.

A long sequence has to be used to illustrate this issue, but that is because we are focused on DNA sequences. This issue is almost guaranteed to show up in every protein sequence alignment. The query sequence is a piece of 4124 bp sequence that is 100% identical to the part from the *M. xanthus* genome on the contig526. The key point here is to show that BLAST is not designed for the kind of alignment we are seeking. This output is in three blocks. It is absolutely not clear why the gap should exist between the first and the second blocks of alignment. Repeated tests show that this gap appears only about 30% of the time. Then the third block is in the reverse direction. After all, no matter what the reason was for

the breakage to be introduced between blocks one and two, BLAST is not suitable for our purpose.

Another problem is that sometimes it simply cannot align correctly. In the following example, the query is 100% identical with the subject in GenBank database, but the algorithm implemented in the BLAST program fails to find the correct alignment. By the way, this problem is due to the special property of the region towards the end of the sequence. A small shift can create some alternative alignment locally, albeit reduces the overall quality of the alignment. Well, that's what BLAST stands for: Basic Local Alignment Search Tool.

```
>gnl|TIGR_246197|contig:521:m_xanthus Myxococcus xanthus
DK 1622 unfinished fragment of genome
Length = 823581

Score = 2677 bits (1392), Expect = 0.0
Identities = 1482/1518 (97%), Gaps = 36/1518 (2%)
Strand = Plus / Plus

Query: 1      ctgctcggcgcgcgagcgctccttgccgcgacggcggttcgatctccagacgcagctcttc 60
            |||
Sbjct: 660201 ctgctcggcgcgcgagcgctccttgccgcgacggcggttcgatctccagacgcagctcttc 660260

Query: 61      ggccctggcgggcttcttcggcctggcgggctccgcctcgacgcgcgcgacctcggcgag 120
            |||
Sbjct: 660261 ggccctggcgggcttcttcggcctggcgggctccgcctcgacgcgcgcgacctcggcgag 660320

Query: 121     acgggctcctcccgcgccagatctcctccgcgcggcggttctctcgcaaggcggggc 180
            |||
Sbjct: 660321 acgggctcctcccgcgccagatctcctccgcgcggcggttctctcgcaaggcggggc 660380

Query: 181     ggactccgcgagacggcgctcctcgatgaggcggggttcttcggcaaggcggttcgcctt 240
            |||
Sbjct: 660381 ggactccgcgagacggcgctcctcgatgaggcggggttcttcggcaaggcggttcgcctt 660440

Query: 241     gagtcgagcctcttcgcgagggcgggcctcctccgaagcgctccgcctcttcgccag 300
            |||
Sbjct: 660441 gagtcgagcctcttcgcgagggcgggcctcctccgaagcgctccgcctcttcgccag 660500

Query: 301     gcgttctgcctcgagccgagcgccctccgcacgctgcgcctcttcgcgaaggcgcgctc 360
            |||
Sbjct: 660501 gcgttctgcctcgagccgagcgccctccgcacgctgcgcctcttcgcgaaggcgcgctc 660560

Query: 361     ttcttcgagcgagcctcttcgcgagccgagcctcttcgcgagggcgcgctcctcttc 420
            |||
Sbjct: 660561 ttcttcgagcgagcctcttcgcgagccgagcctcttcgcgagggcgcgctcctcttc 660620

Query: 421     gaggcgcgcttcttcggcgagacgacgctcttcttcgagccgcgcttcttcggcgagacg 480
            |||
Sbjct: 660621 gaggcgcgcttcttcggcgagacgacgcttcttctcgagccgcgcttcttcggcgagacg 660680

Query: 481     acgctcttcttcgagccgcgcttcttcggctacgcgggcatcttcgcgtaggcgggcgctc 540
            |||
Sbjct: 660681 acgctcttcttcgagccgcgcttcttcggctacgcgggcatcttcgcgtaggcgggcgctc 660740

Query: 541     ctcttcgagcgctcctcttcgcgacgacgctcttctgcgagccgagcctcttctgc 600
            |||
Sbjct: 660741 ctcttcgagcgctcctcttcgcgacgacgctcttctgcgagccgagcctcttctgc 660800

Query: 601     aaggcgcgcttcttcttcagcccgagcctcttcggccagacgacgcttcttctcgagccg 660
            |||
Sbjct: 660801 aaggcgcgcttcttcttcagcccgagcctcttcggccagacgacgcttcttctcgagccg 660860

Query: 661     cgcttcttcgccagccgagcctcttcagccagcgagcttcttcggcgagggcgcgccgc 720
            |||
Sbjct: 660861 cgcttcttcgccagccgagcctcttcagccagcgagcttcttcggcgagggcgcgccgc 660920

Query: 721     ctcaagtagagccgcttctgccagacgagcttcttcgccagacgctccgcctctagccg 780
            |||
Sbjct: 660921 ctcaagtagagccgcttctgccagacgagcttcttcgccagacgctccgcctctagccg 660980
```



```

Query: 781      ggctctttccgcgagccgacgctctcttcaaggcgcttctctgccagccgcgctc 840
Sbjct: 660981  ggctctttccgcgagccgacgctctcttcaaggcgcttctctgccagccgcgctc 661040

Query: 841      ttccgcgagacgaacctctctccgcagacgacgctcttcttcaaggcgcttcttctc 900
Sbjct: 661041  ttccgcgagacgaacctctctccgcagacgacgctcttcttcaaggcgcttcttctc 661100

Query: 901      aagaacgcgctcttccgctaggcggtgctcttcttcaagccgcgctcttctcgagacg 960
Sbjct: 661101  aagaacgcgctcttccgctaggcggtgctcttcttcaagccgcgctcttctcgagacg 661160

Query: 961      agcctcttccgccaacgcgatgctcttcttccgagccgcgcttcttccgccaacgcgctc 1020
Sbjct: 661161  agcctcttccgccaacgcgatgctcttcttccgagccgcgcttcttccgccaacgcgctc 661220

Query: 1021     ttccgcgagacgcgctcttccgctacgcgagcctcttctcgagacgtgctcttccgc 1080
Sbjct: 661221  ttccgcgagacgcgctcttccgctacgcgagcctcttctcgagacgtgctcttccgc 661280

Query: 1081     cagacggcgctcttcttccgagcgtgctcttccagcagacgtgctcttccgccaacgc 1140
Sbjct: 661281  cagacggcgctcttcttccgagcgtgctcttccagcagacgtgctcttccgccaacgc 661340

Query: 1141     acgctcttctt-----cga-----ga--cgagccttctccgagtcgagcctc 1182
Sbjct: 661341  acgctcttcttccgagacgacgctcttccgagtcgagccttctccgagtcgagcctc 661400

Query: 1183     ttccgcgagtcgagcctcttccgcgagacgacgctcttccgccaacgcgagcgtcttctc 1242
Sbjct: 661401  ttccgcgagtcgagcctcttccgcgagacgacgctcttccgccaacgcgagcgtcttctc 661462

Query: 1243     gagacgagcctcttccgccaacgcgagcctcttccgccaacgcgagcctcttccgccaacgc 1302
Sbjct: 661443  gagacgagcctcttccgccaacgcgagcctcttccgccaacgcgagcctcttccgccaacgc 661502

Query: 1303     agcctcttccgccaacgcgagcctcttccgccaacgcgagcctcttcttccgagacgagcttc 1362
Sbjct: 661503  agcctcttccgccaacgcgagcctcttccgccaacgcgagcctcttcttccgagacgagcttc 661562

Query: 1363     ttccgcgagccgagcctcttccgccaacgcgagcctcttcttccgagacgagcctcttccgc 1422
Sbjct: 661563  ttccgcgagccgagcctcttccgccaacgcgagcctcttcttccgagacgagcctcttccgc 661622

Query: 1423     caagcggcgctcttccgagagcgagccttcttccgagagcggcgctcttccgagagcg 1482
Sbjct: 661623  caagcggcgctcttccgagagcgagccttcttccgagagcggcgctcttccgagagcg 661682

Query: 1483     agcttcttccgagcggcg 1500
Sbjct: 661683  agcttcttccgagcggcg 661700

```

Figure. 2.3 Showing the occasional “unexpected” misalignment in the BLASTN output when two 100% identical sequences are aligned. In reality, BLASTN starts local alignment from many fragments in the query sequence. Once those fragments found homology in the database, the alignment is kept and joined together to make a contiguous alignment output. If a region of the sequence has a “good” alternative alignment, the alternative might be kept in the output.

A more serious problem arises when mapping a self-cloned insertion sequence. When a transposon carries an *ori* and an antibiotic marker, it is called self-cloning vector, because one needs only digest the transposon with some restriction enzyme outside the transposon and religate the digested DNA to transform an appropriate host. Nowadays this technique is used widely due to its convenience. A problem with this technique is the fragment cloned in between the two transposon ends is a piece of reshuffled sequence (see Fig. 2.9 and 2.13). The BLAST output will give you two pieces of match with one piece in the reverse direction (Figure 2.4).

```

>gnl|TIGR_246197|contig:526:m_xanthus Myxococcus xanthus DK 1622 unfinished fragment of genome
      Length = 3581774

Score = 535 bits (278), Expect = e-152
Identities = 283/285 (99%), Gaps = 2/285 (0%)
Strand = Plus / Plus

Query: 15      tacgcggagatgctggtcctcaacgccagcacgccggacgatgcggtggcggacgtggcg 74
      |||
Sbjct: 2215129 tacgcggagatgctggtcctcaacgccagcacgccggacgatgcggtggcggacgtggcg 2215188

Query: 75      aaggactgcagcccgaagctggcggagatcatcgggcagaaccagctccgtctgtgtgcgc 134
      |||
Sbjct: 2215189 aaggactgcagcccgaagctggcggagatcatcgggcagaaccagctccgtctgtgtgcgc 2215248

Query: 135     cacgaggacatcctccgcaacctctgcgccaacgcgagcgcgcggtgtcgtctgtgcgc 194
      |||
Sbjct: 2215249 cacgaggacatcctccgcaacctctgcgccaacgcgagcgcgcggtgtcgtctgtgcgc 2215308

Query: 195     aacgtctgcgacttcgcggtgcgcagcgggtgacgctgatggacgtgccgcagatgaag 254
      |||
Sbjct: 2215309 aacgtctgcgacttcgcggtgcgcagcgggtgacgctgatggacgtgccgcagatgaag 2215368

Query: 255     gccgcgcgcgtgcgcgtcttcggccccgaggccgcgg-gggetgcc 298
      |||
Sbjct: 2215369 gccgcgcgcgtgcgcgtcttcggccccgaggccgcggaggc-gcc 2215412

Score = 204 bits (106), Expect = 1e-52
Identities = 107/108 (99%)
Strand = Plus / Plus

Query: 284     ggccgcgggggtgcggacaggattctgggtccgcgctgcgtgacgaanaggtacagcc 343
      |||
Sbjct: 2215023 ggccgcgggggtgcggacaggattctgggtccgcgctgcgtgacgaanaggtacagcc 2215082

Query: 344     gcaggtgctcggcttctcctcctgggctcgtgaagggaacgaggccta 391
      |||
Sbjct: 2215083 gcaggtgctcggcttctcctcctgggctcgtgaagggaacgaggccta 2215130

```

Figure 2.4 It is a bit confusing, at least at first glance at the alignment, where the insertion is on the contig526.

Compare with an intuitive presentation to see what our program can provide in the figure below (Fig. 2.5). The explanation will come later (see Fig. 2.11). But even without an explanation it is quite clear to guess where the insertion is at.

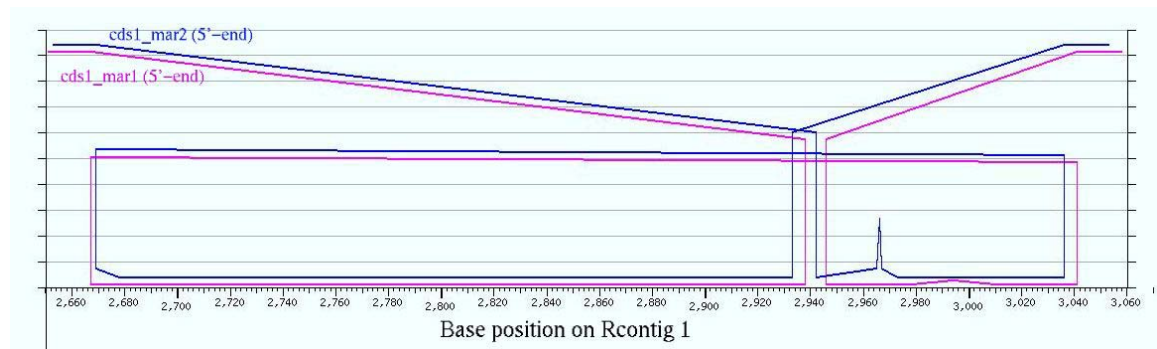


Figure 2.5 Primer mar1 and primer mar2 on template plasmid cds1 both extended through the whole cloned genomic region. The insertion is at the base position 2940. See Fig. 2.11 for detailed explanation.

There are a number of programs for global sequence alignments, such as Clustal and its derivatives. If the query sequences are oriented correctly, it could be the best DNA alignment tool in existence. The problem is that it requires heavy human intervention. An example of Clustal W alignment follows.

```

CLUSTAL W (1.81) multiple sequence alignment

cds40_mar2      -----
contigl8        CCACCAAGGACGCCGCCGATCGCCGGGCTCAGTGTCTCTGCGCATCATCAACGAGCCCA
cds40_mar1      -----

cds40_mar2      -----
contigl8        CCGCCCGCGCCCTGGCCTACGGCTTGACAAGGTGCAGGACGCTGGCACCAGCGCATCG
cds40_mar1      -----

cds40_mar2      -----
contigl8        CCGTCTACGACCTGGGCGCGGCACCTTCGATATCTCCATCCTGGAGCTGAACGCCGCGG
cds40_mar1      -----

cds40_mar2      -----
contigl8        TGTTCGAAGTGAAGAGACCAACGGCGACACGTTCTTGGGCGCGAGGACTTCGACCAGC
cds40_mar1      -----

cds40_mar2      -----
contigl8        GCCTCATCGACTACCTGGCCAAGCGCTTCGCGGAATCCAACAACGGGCTGGACCTGCGCA
cds40_mar1      -----

cds40_mar2      -----
contigl8        AGGACCGCATGGCGCTGCAGCGCCTGAAGGAAGCGCCGAGCGCGCCAAGCACGAGCTGT
cds40_mar1      -----

cds40_mar2      -----
contigl8        CCAGCGCGCCCGAGACGGAGGTGAACCTGCCGTTTCATCACCGCGATGCCCTCCGTCCCA
cds40_mar1      -----

cds40_mar2      -----
contigl8        AGCACCTCACGGAGACCGTGGACCGCGCACCTTCGAGGCGCTGCTGACGACCTCATCG
cds40_mar1      -----

cds40_mar2      -----
contigl8        ACCGACCATCGAGCCGTGCCGATTGCCCTGAAGGACCGGGCATTCCCGCGCAGCAGA
cds40_mar1      -----

cds40_mar2      -----
contigl8        TCAACCAGGTGCTGCTGGTGGGCGCATGACGCGCATGCCGCGCTGCAGCAGAAGTGA
cds40_mar1      -----

cds40_mar2      -----
contigl8        AGGAGTTCTTCGGCAGGGAGCCTCACAAGGGCATCAACCCGGACGAGGTGCTGCCGTGG
cds40_mar1      -----

cds40_mar2      -----
contigl8        GCGCGGCCATCCAGGGCGGTGTGCTCAAGGCGGAGGTGAAGGACGTCCTCTGCTGGACG
cds40_mar1      -----

cds40_mar2      -----
contigl8        TGACGCCGCTGTGCTCGGTGTGAGACGGCGCGGTGTCTTCACGAAATCATCGACA
cds40_mar1      -----

cds40_mar2      -----
contigl8        AGAACACCACCATCCCCTGCAAGAAGAGCCAGTGTCTCCACCGCGTGGACAACCAGC
cds40_mar1      -----

cds40_mar2      -----
contigl8        CGCTGGTGAGCGTGCACGTGCTCCAGGGCGAGCGTGAGATGGCGCGGACAACAAGACGC
cds40_mar1      -----

cds40_mar2      -----
contigl8        TGGCGCGCTTCGAACTGGTGGGCAATCCCCCGGCCCGCGCGCGCTGCCGAAATCGAGG
cds40_mar1      -----

cds40_mar2      -----
contigl8        TGTCTGTCGACATCGACGCCAACGGCATCGTCCACGTCAGCGCCAAGGACCTGGGCACCG
cds40_mar1      -----NCATTATCAGCCAAACCTGTTACTCGCTCCATG

cds40_mar2      -----
contigl8        GCAAGGTTCAGCAGGTGCGCGTGGTGAGCAACTCCGGCCTGTCCGAGGCGGAAATCCAGG
cds40_mar1      CCTTCGTTCTCCGATCGNAGCTAGATAGAGAGATCGTCGGCGTATGTTGCACAGCTCCGC

cds40_mar2      -----
contigl8        CGATGATTTCCGACGCCCACTCGACGCTCCGACGACAAGAAGAAGGAGCTGGCGG
cds40_mar1      CAACTCCTTCTTCTTGTACGTGAGGCGTGCAGCTGGGCGTCCGAAATCATCGCCT

cds40_mar2      -----
contigl8        AGCTGCGCAACAACGCCGACGGCCTCATCTACACGCGGAGAAGAGC-CTGGAGGAGTAC
cds40_mar1      GGATTTCCGACCTCATAAGGCACGGAGTTGCTC-ACCACGCGCACCGCTGAACCTTGC
* * *

cds40_mar2      GCGAGCCTCCTGTC---GGAGAAGGACCGCGAGGAAATCAAGGCGGACCTGGAGCGCCTC
contigl8        GCGAGCCTCCTGTC---GGAGAAGGACCGCGAGGAAATCAAGGCGGACCTGGAGCGCCTC
cds40_mar1      CGCGGCCCTACGTCCTTGGCGCTGAACCTGGACGATGCCCATGGCGTCCATGTCGAACGAC
* * * * *

```

```

cds40_mar2      AAGGAGGTGCTCAACACCTCCGACGCGCGGTGCTCAAGGAATCCTTCAGCGCCTGGAA
contigl8        AAGGAGGTGCTCAACACCTCCGACGCGCGGTGCTCAAGGAATCCTTCAGCGCCTGGAA
cds40_mar1      ACCTCGAATTGCGGCACGACCGCT---CGGCGCTGCGGAATGC--CCACCATTTGAA
*   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *

cds40_mar2      G-GCAGCGCTACCGCATCGCGGACGCCATCT---ACACGGGCCAGGCGAGCTGAACGCT
contigl8        G-GCAGCGCTACCGCATCGCGGACGCCATCT---ACACGGGCCAGGCGAGCTGAACGCT
cds40_mar1      GCGCGCCAGCCTCTTTGTTGTCGCGCCCATCTCACGCTCGCCCTGGAGCACGTGCACGCT
*   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *

cds40_mar2      CGCAATCGCCTCCGCGCTC---CAGCGTAGACTGCCTGCCGCGCAGTCAGTCCCCCTGG
contigl8        CGCAATCGCCTCCGCGCTC---CAGCGTAGACTGCCTGCCGCGCAGTCAGTCCCCCTGG
cds40_mar1      CACCAGCGGCTGGTTGTCCACGCGGTGGAGAACACTGNCCTCTCTTTCACGAAATAAT
*   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *

cds40_mar2      AGCGCATACATGGA-CGTACCG--AAGGCATCGTCATCTCC-CTCATACCG--CCATG
contigl8        AGCGCATACATGGA-CGTACCG--AAGGCATCGTCATCTCC-CTCATACCG--CCATG
cds40_mar1      GGCCTTCTGTCAATCATTTTCGTAAACACACCGCTGCTCTTAAACACCGAGCGACG
*   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *

cds40_mar2      GTGGTGGGGGTGCCCTGTTCTGTCTCACGCTTCGCTTCTCCCTCAGGCCCTGGTG---
contigl8        GTGGTGGGGGTGCCCTGTTCTGGGCTCACGCTTCGCTTCTCCCTCAAGCCCTGGTG---
cds40_mar1      GCGACGNTCCGCGGTACCGAGAAGAGAGAACGCTCTCCCTTAAACGACACCAACCC
*   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *

cds40_mar2      -GAAGCCTTCTCCCGCTGAAGTAAACCCAGCACGGGGCATG---GACCTGCGGTTCTC
contigl8        -GAAGCCTTCTCCCGCTGAAGTAAACCCAGCACGGCGCATG---GAGGTGCGGCTG-C
cds40_mar1      TGAATTTCCCAACACAGTCACAACTCAATAGGTTTCAGGCGCGCTCAGGTGCC
*   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *

cds40_mar2      T--CCGCGAACGCATCGCGCACCTGTAACAGTGTGGAAGGACACGGCTCGTGTATGA
contigl8        T--CCGCGAGCGCATCGCGCACCTGGAGCGGCTGTGGAAGGACACGGCTCATGGATGA
cds40_mar1      TTACCGCATACGGGGGNCACCGACGCGCGCNCACCGTCCCGGAGAGACGNAGC
*   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *

cds40_mar2      CAGCCTGTCCC-TGGCGCTTCTACCGGCTCTTCTGAATGCCGCGG-GCACCTAGGTG
contigl8        CCACCTGTCCC-TGTCCCTT-----
cds40_mar1      CAGTCCAGCCCATNTCGCTTTTACCTCTGTCAAAANCGNGGGGCGGAGCGCCCGAACA
*   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *

cds40_mar2      GCTACGGCGATGCGTAACCGCTCTAATCGTCTAGCCGATTAGGGGTACGGAAGGC
contigl8        GAAACGCGACCGCGGAAAAACCACTTACCACCCGATCGGCGTACATAGGAAACGGAAAC
cds40_mar1      GGGGCTTCCCCATTATCGCCCTCCGCCCCTGTCTCTCCCCCCCCCCTGGTAGCAGAG
*   *   *   *   *   *   *   *   *   *   *   *   *   *   *   *

cds40_mar2      GGGACACGCGGCGCACGCGCC-----
contigl8        -----
cds40_mar1      -----

cds40_mar2      AGGTCGGTGGTTGCTGGCGGGCTTGGCGTCCCCGATATGTAATTCGCTGTGTCGTGTC
contigl8        -----
cds40_mar1      -----

cds40_mar2      CCTTCCCTTGGCGCGCTCCCTGTCTCGCGACCTTCTGTCGCGGCTGCTTTGTCTGAT
contigl8        -----
cds40_mar1      -----

cds40_mar2      CGTTCGCGTCACCTTCTTGTAGTGGCGGTGGGTGGCGGCTGTGAGCC
contigl8        -----
cds40_mar1      -----

```

Figure 2.6. Clustal W output for multiple alignment of cds40_mar1, cds40_mar2 with contig 18. Note that cds40_mar1 and cds40_mar2 are the flanking sequence of transposon insertion cds40. Therefore cds40_mar1 and cds40_mar2 are not supposed to overlap, and should extend in the opposite direction from the insertion point. The fact that cds40_mar2 and contig 18 have a long stretch of very good homology indicates their alignment is valid. The cds40_mar1 has only sporadically alignment suggests the alignment is questionable. In fact, Clustal does not care about how similar the given sequences are, it will output an alignment for the given sequences. Human intervention is absolutely required.

The above example shows a problem. The cds40_mar1 and cds40_mar2 essentially mapped to the same region on contig 18. We expect them to map tandem to a region in opposite (head to head) direction, like the one in Figure 2.7 below. The problem arises because the cds40_mar1 should be converted into its complement strand before sent to Clustal. Clustal aligns sequences as they are provided. It doesn't care whether the complement of a sequence

match the target sequence any better than the original sequence. Once cds40_mar1 is converted to its complement, the alignment produces the desired alignment (Figure 2.7). I made this an automatic process in my program.

```

CLUSTAL W (1.81) multiple sequence alignment

cds40_mar2 -----
contigl8  CCACCAAGGACGCCGGCGCATCGCCGGCTCAGTGTCTGCGCATCATCAACGAGCCCA
cds40_mar1 -----

cds40_mar2 -----
contigl8  CCGCCGCGGCCCTGGCCTACGGCTTGGACAAGTGCAGGACGGTGGCACCGAGCGCATCG
cds40_mar1 -----

cds40_mar2 -----
contigl8  CCGTCTACGACCTGGGCGCGGCACCTTCGATATCTCCATCCTGGAGCTGAACGCCGGCG
cds40_mar1 -----

cds40_mar2 -----
contigl8  TGTTCGAAGTGAAGAGCACCAACGGCGCACGTTCTTGGGCGCGAGGACTTCGACCAGC
cds40_mar1 -----

cds40_mar2 -----
contigl8  GCCTCATCGACTACCTGGCCAAGCGCTTCGCGGAATCCAACAACGGGCTGGACCTGCGCA
cds40_mar1 -----

cds40_mar2 -----
contigl8  AGGACCGCATGGCGCTGCAGCGCTGAAGGAAGCGCCGAGCGCGCAAGCACGAGCTGT
cds40_mar1 -----

cds40_mar2 -----
contigl8  CCAGCGCGCCCGAGACGGAGGTGAACCTGCGCTTCATCACCGCGATGCTCCGTCCTCA
cds40_mar1 -----GGCGCGGTGCGCGG

cds40_mar2 -----
contigl8  AGCACCTCACGGAGACCGTGGACCGCGCCACCTTCGAGGCGCTGGTGCAGGACCTCATCG
cds40_mar1 -----CGTGTCCCGTTTCCGTTTCCCTAT-GTACGCCGATCGGGGTGGTAA-GTGGGTTTTCGCG

cds40_mar2 -----
contigl8  ACCGCACCATCGAGCGCTGCCGGATTGCGCTGAAGGACGCGGCGATTCCCGCGCAGCAGA
cds40_mar1 -----GTCGCGTT-TCTGTTCGGGGCGCTCCGCCCCNCGN-TTTTGACAGGTAAGAAGCANA

cds40_mar2 -----
contigl8  TCAACAGGTGCTGCTGGTGGGCGCATGACGCGCATGCCGCGCTGCAGCAGAAGGTGA
cds40_mar1 -----TGGGCTGGACTGGCTNCGTCTCTCCCGGACG--GTGNCGCGCTGCTGCTGGTGNCC

cds40_mar2 -----
contigl8  AGGAGTTCTTCGGCAGGAGCGCTCACAAGGG--CATCAACCGGACGAGGTCTGCGCGT
cds40_mar1 -----CCCCGTATCGGTAAGGCAGCGCTGACGGCGCGCTGAACCCCTATTGAGGTGTGACTGT

cds40_mar2 -----
contigl8  GGGCGCGCCATCCAGGGCGGTG--TGCTCAAGGGCGAGGTGAAG--GACGTCTCTCTGC
cds40_mar1 -----TGGTGGGGAATTCAAGGTGTTGTGCGTTAAGGAGAGGCGTTCTCNTCTCTCGGTAC

cds40_mar2 -----
contigl8  TGGACGTGACGCGCTGTGCTCGGTGTGCGAGACG-GCCGCGGTGTCTTCACGAAATC
cds40_mar1 -----CGCGGANCCGTCGCGCTGCTGTTAAGAGGACGAGCGGTGTGTTTACGAAATG

cds40_mar2 -----
contigl8  ATGACAAGAACAACCATCCCTGCAAGAAGAGCCAGGTGTTCTCACCGCGCTGGAC
cds40_mar1 -----ATTGACAAGAAGCCATTATTTCGTGCAAGAAGAGNCAGGTGTTCTCACCGCGCTGGAC

cds40_mar2 -----
contigl8  AACCAGCGCTGCTGAGCGTGACGTGCTCAGGGCGAGCGTGAGATGGCGCGGACAAAC
cds40_mar1 -----AACCAGCGCTGCTGAGCGTGACGTGCTCAGGGCGAGCGTGAGATGGCGCGGACAAAC

cds40_mar2 -----
contigl8  AAGACGCTGGCGGCTTCGAACTGGTGGGCTTCCCGCGCGCGCGG-CGTGCCGCA
cds40_mar1 -----AAGAGGCTGGCGGCTTCGAAATGGTGGGCTTCCGAGGCGCGGAGCGGTGTGCCGCA

cds40_mar2 -----
contigl8  AATCGAGGTGTGCTTCGACATCGACGCCAACGGCATCGTCCAGCTCAGCGCCAAGGACCT
cds40_mar1 -----ATTGAGGTGTGCTTCGACATGGACGCCATGGGCATCGTCCAGTTCAGCGCCAAGGACGT

cds40_mar2 -----
contigl8  -GGGCACCGGCAAGGTTACGAGGTGCGCGTGGTGGCAACTCCG-GCCTGTCCGAGG-C
cds40_mar1 -----AGGGCGCGGCAAGGTTACGAGGTGCGCGTGGTGGCAACTCCGCTGTATGAGGTC

cds40_mar2 -----
contigl8  GGAAATCCAGGCGATGATTTCCGACGCCAGTCGACGCTCCGACG-ACAAGAAGAAGA
cds40_mar1 -----GGAAATCCAGGCGATGATTTCCGACGCCAGTCGACGCTCCGACGTACAAGAAGAAGA

cds40_mar2 -----
contigl8  AGGAGCTGGCGGAGCTGCGCAACA-ACGCCGAGCGCTCATCTACACGACGGAGA--AGA
cds40_mar1 -----AGGAGTTGGCGGAGCTGTGCAACATACGCCGACGATCTC-TCTATCTAGTNCGATCGGA
* *

cds40_mar2 -----
contigl8  TCGCCAACCTGTTACGCGAGCCTCCTGTCGGAGAAGGACCGCGAGGAAATCAAGCGGAC
cds40_mar1 -----GC-CTGGAGGAGTACGCGAGCCTCTGTCTGGAGAAGGACCGCGAGGAAATCAAGCGGAC
* * * * * * * * * * * * * * * * * * * * * *

cds40_mar2 -----
contigl8  CTGGAGCGCCTCAAGGAGGTGCTCAACACCTCCGACGCGCGGTGCTCAAGGAATCCTTC
cds40_mar1 -----CTGGAGCGCCTCAAGGAGGTGCTCAACACCTCCGACGCGCGGTGCTCAAGGAATCCTTC

cds40_mar2 -----
contigl8  CAGCGCCTGGAAGGCAGCGCTACCGCATCGCGGACGCCATCTACAGGGCCAGCGAGC
cds40_mar1 -----CAGCGCCTGGAAGGCAGCGCTACCGCATCGCGGACGCCATCTACAGGGCCAGCGAGC

```

```

cds40_mar2      TGAACGCTCGCAATCGCCTCCGCGCTCCAGCGTAGACTGCCTGCCGCGCAGTCAGTCCCC
contigl8        TGAACGCTCGCAATCGCCTCCGCGCTCCAGCGTAGACTGCCTGCCGCGCAGTCAGTCCCC
cds40_mar1      -----
cds40_mar2      CTGGAGCGCATACATGGACGTACCGAAGGCATCGTCATCTCCCTCATCACCGCCATGGT
contigl8        CTGGAGCGCATACATGGACGTACCGAAGGCATCGTCATCTCCCTCATCACCGCCATGGT
cds40_mar1      -----
cds40_mar2      GGTGGGGGTGCCCTGTTCTGTCTCAGCTTCGCTTCTCCCTCAGGCCCTGGTGAAGC
contigl8        GGTGGGGGTGCCCTGTTCTGGGGTCACGCTTCGCTTCTCCCTCAAGCCCTGGTGAAGC
cds40_mar1      -----
cds40_mar2      CTTCTCCGGCTGAAGTAAACCCAGCACGGGGCATGGACGTGCGGTTGTCTCCGCGAAC
contigl8        CTTCTCCGGCTGAAGGAAACCCAGCACGGGGCATGGAGGTGCGGCTG-CTCCGCGAGC
cds40_mar1      -----
cds40_mar2      GCATCGCGCACCTGTAACACGTGCTGGAAGGACACGGCCTCGTGTATGACAGCCTGTCCC
contigl8        GCATCGCGCACCTGGAGCGCGTGTGGAAGGCGACGGCCTCATGGATGACCACCTGTCCC
cds40_mar1      -----
cds40_mar2      TGGCGCTTCTACCGGCTCTTCTGAATGCCGGCGCACCTAGGTGCGCTACGGCGATGC
contigl8        TGTCCCTT-----
cds40_mar1      -----
cds40_mar2      GTAACCCGCTCTAATCGTCTAGCCGGATTAGGGGTACGGCAAGCGGGGCTTCCCCCA
contigl8        -----
cds40_mar1      -----
cds40_mar2      TTATCGCCCTCCGCCCTGTCTCTCCCCCCCCCTGGTAGCAGAGGTCGGTGGTTG
contigl8        -----
cds40_mar1      -----
cds40_mar2      CTGGCGGGCTTGCGGTCCCCGATATGATTTCCGTGTGTCGTGCCCTTCCCTTGCGC
contigl8        -----
cds40_mar1      -----
cds40_mar2      GCGTCCCCTGTCTCGCGACCTTCGTGCCCGTGCTTTGTCTGATCGTTCCCGTCACC
contigl8        -----
cds40_mar1      -----
cds40_mar2      TTTCTTGAGTGGCGGTGGGTGGCGCCTGTGAGCC
contigl8        -----
cds40_mar1      -----

```

Figure 2.7. Clustal W output for multiple alignment of cds40_mar1 (complement), cds40_mar2 with contig 18. cds40_mar1 (complement) and cds40_mar2 are the flanking sequence of transposon insertion cds40. After converting the cds40_mar1 to its complement, cds40_mar1 (complement) and cds40_mar2 do map to the flanking region of the insertion cds40 in the opposite direction to each other from the insertion point. Here one can find the insertion point locates to a narrow region where the three sequences overlap. If some attention is paid to the overlap region, one can find the top two have very good alignment except the first a few bases. Therefore the insertion point should be located at the beginning of the overlap or some bases upstream from there where the homology between the cds40_mar1 (complement) doesn't have good homology with the contig 18.

It is obvious that even the best possible combinations of existing programs to map insertion point needs an additional program to convert the DNA sequence into its complement. Even when the complement sequence generating program is available, this approach is quite clumsy as can be seen in the above. A researcher has to find out whether the complement of a sequence should be used for the alignment. Another problem may occur if one has many insertions to be mapped. One has to go through the same procedure over and over again. In this situation, one more problem is how does the researcher know which insertion sequence is supposed to be aligned with which contig. Or how many insertion sequences are going to

be mapped to a particular contig. Manually pairing each and every insertion point sequence with every contig is not a very efficient way to find the correct matches. A more efficient way to map the insertion points is obviously of value, especially when the number of insertions screened is big. However, there is no such program available as far as I know. Maybe this is because biologists are not complaining enough about the chore of mapping insertion sites, or because the fact that genomic sequences are not known until recently, and to be able to map an insertion site to a precise genomic location is a new development.

STATEMENT OF THE PROBLEM

In molecular biology, it is a common practice to generate mutants to study the effects of a particular gene. One of the most popular ways of generating mutants is by inserting a piece of DNA into the genomic copy of the gene of interest to disrupt the gene. Transposon insertion is one of the most used methods for this purpose. However, transposon insertion is a random process. That is, transposons can occur at a large number of sites in a genome. Therefore to find out the location of the insertion on the genome requires sequencing the genome region flanking the insertion, then putting the flanking region into a broader context of genomic sequence to identify the location of the insertion, and the gene(s) the insertion interrupts. This chapter discusses the first computer program specifically designed for identifying the location of the (transposon) insertion and presenting the result in a picture showing the base coordinates in a relevant contig of the genome.

This chapter is focused on designing, developing, and demonstrating a prototype utility for mapping insertion points to their genomic sequence positions automatically.

TRANSPOSON A transposon is a piece of DNA that is able to move from one genomic location to another. Naturally occurring transposons carry a gene encoding a transposase that is critical for the translocation of the transposon. The transposon *mariner* used in this study has been heavily engineered to facilitate the creation of insertions and the recovery of the genes carrying the insertion mutation. It retains only the two termini required for transposition. Inside the two termini are an *ori* sequence and a kanamycin resistance gene. This version of *mariner* is called *magellan-4*. The presence of an *ori* allows the piece of DNA to be replicated independently of the genome when the transposon is cloned into *Escherichia coli*. The kanamycin gene makes the presence of the insertion selectable by the antibiotic kanamycin.

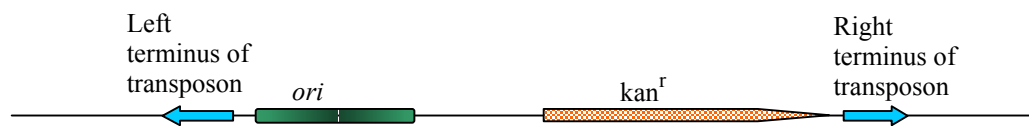


Figure 2.8 The structure of the transposon, *magellan-4*.

This transposon is said to be self-cloning because the presence of the *ori*. That is, if the transposon is taken out of the genome (e.g. via restriction digestion) and then two ends are ligated, it can propagate in appropriate bacterial host such as *E. coli*. This feature makes cloning more convenient, and more efficient.

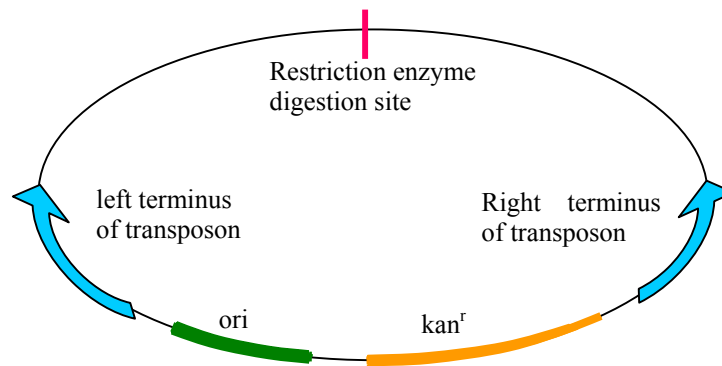


Figure 2.9 When the transposon is removed from a genome and the two ends are joined with each other; it is self-replicative in appropriate bacteria, in *Escherichia coli* for instance. That is called “self cloning”.

DETERMINATION OF DNA SEQUENCE

Once the insertion point is cloned, its sequence is determined by the sequencing facility at OMRF (Oklahoma Memorial Research Foundation) using dideoxy method (Sanger’s method).

DEFINITION OF ALIGNMENT Given a sequence $A = a_1, a_2, \dots, a_{m-1}, a_m$, of length m , and a sequence $B = b_1, b_2, \dots, b_{n-1}, b_n$, of length n , both consisting of symbols from an alphabet Z of size c . Aligning A and B can be considered as the process of transforming sequence A into B by replacement, insertion, or deletion of symbols in the A (*cf.* Rognes and Seeberg, 1998).

As it is used in processing biological information, alignment implies a meaningful alignment. That is, the presence of an alignment between two sequences means the two

sequences have a significant similarity (homology). Therefore, it is necessary to set a scoring system to separate the meaningful alignments from the rest. The stringency of the scoring system is completely artificial. Usually it is set by trial-and-error and / or preexisting data.

Based on this definition, the program described here is designed to symbolically align and present the homology without presenting any base codes. The purpose of this program is to align the insertion point flanking sequence with the native genomic sequence to find and present the insertion point.

DEFINITION OF SEQUENCE MAPPING Given a sequence of $S = s_1, s_2, s_3, \dots s_{m-1}, s_m$ of length m , the complement of S if there is one $C = c_1, c_2, c_3, \dots c_{m-1}, c_m$ of length m , and a sequence $T = t_1, t_2, t_3, \dots t_{n-1}, t_n$ of length n , all consisting of symbols from alphabet Z of size c . Mapping S to T can be considered as the process of finding regions of alignments between S and T , and between C and T , and presenting the regions of alignment in the order of sequence S against T .

For visual convenience, the disjunct regions of alignments will be linked with lines to symbolize the flow of the sequence (Figure 2.11). This presentation is particularly useful for biologists who are more interested in the map than in the process of mapping. The sequences used in mapping usually are from sequencing readouts, which are usually several hundred bases long or slightly longer. The target sequences usually are much longer. They may be assembled contiguous fragments from a sequencing project, or from some databases. If so, they are called contigs.

As an example, sequence mapping is used here to locate the insertion point of transposons, therefore also called insertion point mapping.

RESULTS

The DNA mapping algorithm

For a genome size of 10,000,000 bases long (*Myxococcus xanthus* genome size ~9.2 million bases), statistically a fragment of 12 bases is long enough to guarantee its uniqueness, assuming the base sequence is completely random. In reality the sequence fragment has to be longer than 12 bases to guarantee its uniqueness. First, the genome sequence is not a random sequence of bases. Second, it is expected that some repeated sequence fragments exist in the genome. However, it is unlikely that a stretch of sequence a hundred or more bases long is to be found in the genome many times, unless it happens to be one of the repeated sequence fragments. For this reason, to map the insertion points needs only about a hundred bases long, continuous, good homology between the flanking sequences and the contig. Therefore, the program is designed to measure the similarity of a stretch of homologous sequence pairs that is long enough to be unique to eliminate spurious matches.

The flanking sequences from a sequencing machine usually are a few hundred bases long or slightly longer. They are much too long to be directly used for homology distance calculation. They are chopped into query fragments (**F**) of manageable size 12 to 50 bases long. Although it can be longer, it is found from test runs that longer fragments significantly reduce the computation speed under our settings. Using shorter fragments also reduces the computation speed. This fragmentation process is good for sensitivity and flexibility as well, and is called dynamic programming.

To calculate the homology is to compare the query fragment with every piece of subject

sequence of the same length to find out the fewest base changes (by insertion, deletion and/or replacement) needed to make the two identical. This is called calculating the Levenshtein distance (**D**) between the two fragments. Each **D** is collected as an element of a matrix **D_{F1}**. This process starts with comparing a query fragment with the fragment from bases 1 to **F** on the subject sequence, then with the fragment from bases 2 to **F**+1, and so on till the end of the subject sequence is reached. This is called sliding window, a variant of dynamic programming.

$$\mathbf{D}_{F1} = \{ \mathbf{D}_{F1_1}, \mathbf{D}_{F1_2}, \dots \}$$

Then chop the second fragment from the query sequence, and compare it with the subject sequence from the beginning as before. Repeat the process till the end of the query sequence. A set of matrix is collected:

$$\begin{aligned} \mathbf{D}_{F1} &= \{ \mathbf{D}_{F1_1}, \mathbf{D}_{F1_2}, \dots \} \\ \mathbf{D}_{F2} &= \{ \mathbf{D}_{F2_1}, \mathbf{D}_{F2_2}, \dots \} \\ \mathbf{D}_{F3} &= \{ \mathbf{D}_{F3_1}, \mathbf{D}_{F3_2}, \dots \} \\ &\dots \end{aligned}$$

This set of matrix can be combined in to a single two dimensional matrix:

$$\mathbf{D}_F = \{ \mathbf{D}_{F1}, \mathbf{D}_{F2}, \mathbf{D}_{F3}, \dots \}$$

That is the same as:

$$\mathbf{D}_F = \{ \begin{aligned} &\{ \mathbf{D}_{F1_1}, \mathbf{D}_{F1_2}, \dots \} \\ &\{ \mathbf{D}_{F2_1}, \mathbf{D}_{F2_2}, \dots \} \\ &\{ \mathbf{D}_{F3_1}, \mathbf{D}_{F3_2}, \dots \} \\ &\dots \end{aligned} \}$$

This matrix makes it easy to find the homologous regions. Since Levenshtein distance essentially measures the differences between the two sequence fragments, to find homology means to find regions between the query and subject sequences with very small \mathbf{D} . Assume \mathbf{D}_{F12} is the smallest value in the matrix \mathbf{D}_{F1} , that is to say the first fragment from the query sequence has the best match to the subject sequence at a position starting from the base 2 on the subject sequence. If we do the same for every matrix from \mathbf{D}_{F1} to $\mathbf{D}_{F\text{last}}$, we now have every fragment's best match position in another matrix (Figure 2.10).

8	4	5	7	9	7	8	8	6	5	6	8	7	9	8	8	7	6	8	6	5	8	7	5	4	8	6	8	7	9	7	8	6	9	7	8	6	6	7	8	6	9	8	7
8	6	8	7	6	9	8	7	5	4	2	1	3	4	5	7	6	8	9	6	8	7	7	8	6	7	8	7	8	7	7	8	7	6	9	8	8	7	6	5	7	8	8	6
6	7	6	8	5	8	8	7	6	9	8	6	7	5	7	6	8	7	7	5	3	0	2	4	7	5	3	7	9	8	8	6	7	5	8	7	7	6	9	8	6	8	7	6
7	9	7	8	6	7	9	5	9	7	8	6	9	6	8	9	6	9	8	7	8	6	8	7	9	7	6	8	8	6	3	0	2	6	8	7	8	6	7	8	8	7	9	7
8	6	7	6	7	8	7	8	5	7	8	6	7	8	9	8	7	8	8	9	5	6	7	6	5	8	7	4	6	8	4	8	6	7	8	5	9	6	4	3	1	3	5	7
6	8	7	8	7	8	7	8	8	7	5	8	9	7	8	8	5	3	7	6	8	5	6	3	6	8	7	7	8	8	6	9	8	8	7	6	4	7	8	6	7	8	9	7
8	7	6	8	7	5	8	7	6	7	6	8	6	7	5	7	8	8	7	8	8	9	8	7	8	7	8	9	7	8	8	8	6	9	9	8	7	8	7	8	6	7	5	8

Figure 2.10. An example of 2 dimensional Levenshtein distance matrix. The fragment size is 10 bases long. All the smallest \mathbf{D} s for each fragment are in red. The bold red numbers represent a stretch of extended homology. However, the first fragment is 4, therefore, unlikely to be truly homologous, and is removed by applying \mathbf{L} .

Once we know the best matches for each and every fragment from the query sequence, we want to know whether each of those matches can be linked together with an acceptable overall homology. Since the smallest Levenshtein distance does not mean a good homology, a difference between two fragments of 4 out of 10 bases can be the smallest within a matrix, but the difference is too big for it to have any biological bearing. Therefore, those elements (and the matrix they are in) can be removed from further calculation (the line 1 in the Figure 2.10). For this purpose, a cutoff value \mathbf{L} is experimentally set to $0.3\mathbf{F}$ to retain all possible matching elements. If the \mathbf{L} value is set too low, some of the marginal fragments may be excluded from matching.

To make the Levenshtein distance calculation tolerate possible clusters of sequence errors

that might otherwise break the homology, the **D** value is converted to a weighted distance **D_w** (dividing the sum of Levenshtein distances of consecutive query fragments by the numbers of bases in that consecutive stretch of DNA). Mathematically it is defined as follows:

$$\mathbf{D}_w = \left(\sum_{i=m, \mathbf{D}_i < 0.3\mathbf{F}_i}^n \mathbf{D}_i / \sum_{i=m, \mathbf{D}_i < 0.3\mathbf{F}_i}^n \mathbf{F}_i \right)$$

i – count of consecutive fragments of query sequence

m – the number of the first fragment in a consecutive stretch of fragments

n – the number of the last fragment in a consecutive stretch of fragments

For example, if a stretch of homology ends with a fragment half of which is homologous, and the rest is random. The **D** value will be about 0.4F. Therefore it is not going to be part of the homology. But if the end fragment starts with 70% identical bases, the **D** value will be less than 3, then it is likely to be included in the final homology stretch. If the **L** value is set too high the amount of calculation will increase significantly.

This **D_w** allows occasional errors, even some error clusters in the homology, because the **D_w** value is distributed over the whole length of the homology. With **D_w** we can weed out stretches of non-homologous (**D** is much bigger than 0) elements by setting another limit **H**. If **D_w** is under the limit **H** (experimentally set at 0.1), the stretch of homology is retained as biologically meaningful homology (Figure 2.11). It is possible to find several stretches of biologically meaningful homology between a query sequence and a subject sequence, interrupted by pieces of sequences with **D** above 0.3F (Figure 2.15). The retained value (in the example above, 1, 0, 0, and 1 in Figure 2.10) is stored in another matrix for further

processing.

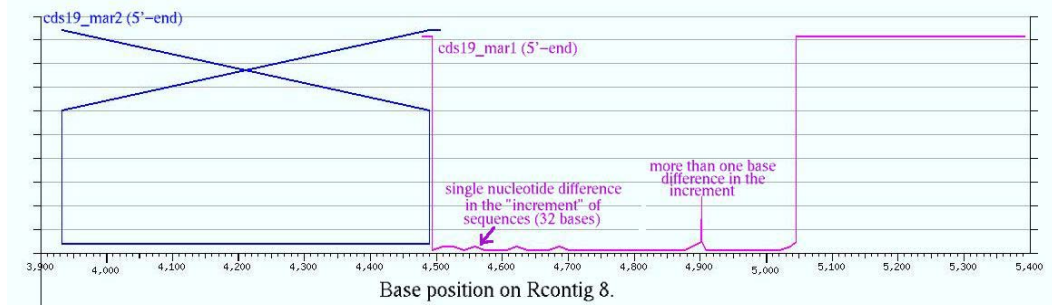


Figure 2.11. Insertion cds19 maps to the Rcontig8, at base position 4490. The primer mar1 is oriented in the same direction as the contig. The primer mar2 is oriented in a reverse direction, therefore the line representing the query sequence derived from mar2 has a cross in the figure.

How to read the mapping result: Take the cds19_mar1 as an example. The little horizontal purplish line (about 20 bp) on the top represents a very small piece of 5'-sequence does not match the Rcontig8. Then the vertical line simply means that starting from that base position (~4490 bp) on the Rcontig8 the two sequences have significant homology. Since the homology is not 100%, the bottom line is not straight for ~200 bp, then the two sequences matched 100% for about 180 bp, then comes another region of disagreement at around 4900 bp position on the Rcontig8. It is followed by a ~100 bp stretch with 100% agreement. Towards the end for about 350 bp, the two sequences do not match any more; therefore, the query sequence is lifted up high. For cds19_mar2 is in the reverse direction, the mar2 has go to the insertion point shown as a diagonal line. Once reached the position, mar2 drops vertically down to the Rcontig8, meaning the mar2 matches the Rcontig8 from its very beginning. Then the mar2 matches the contig completely (100%) from base 4490 to base 3930. At the very end there is about ~20 bp does not match the Rcontig8, and is lifted back up and positioned at the 3'-end of the mar2 line via another diagonal line. In short, the aligned part of each query sequence is the segment that is close to and parallels the X-axis. Different sizes of bends in the aligned segment represent various degrees of errors in those parts of the query sequence.

There is another molecular biology aspect that has to be considered for insertion point mapping. Although the flanking sequences are derived from the primers annealed to the terminus of an insertion (for example, mar1 in Figures 12 and 13), the template is a piece of circularized genomic DNA, depending on the distance from the restriction site to the insertion point, the flanking sequence might extend beyond the distance between the primer

and the restriction site. Therefore the flanking sequence to the other side of the insertion (the mar2 side of the insertion) may be included downstream of the flanking sequence of the primer mar1 side of the insertion (Figures 2.12 and 2.13).

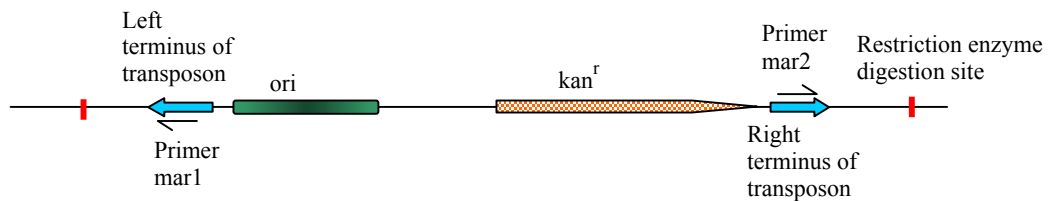


Figure 2.12 The insertion in the genome. Sequences derived from the upstream primers will never reach the sequence flanking the downstream side.

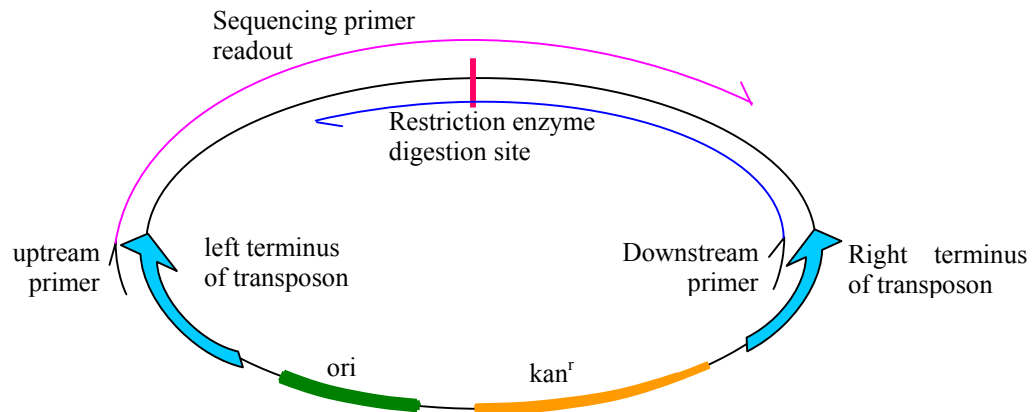


Figure 2.13 The insertion is cloned in a self-cloning plasmid. The sequence derived from the upstream primer mar1 (pink) may extend to cover the region flanking the other side of the insertion (the mar2 side of the insertion).

Note: It doesn't matter from which strand the sequence started, the sequence beyond the restriction site is read from the complementary strand of the genome. Therefore to map this part of the sequence, it is necessary to convert it into the complement.

To make the program to detect and align the complement as well as original sequences, the complement alignment is executed for all query sequence fragments. The mapping process between the complement and the subject sequences follows the same protocol as before, except that each fragment of the query sequence is converted to its complement before the alignment process. In reality, in our real world mapping process, several cases were discovered where the sequence readouts from the sequencing primer extend beyond the cloning site to cover the sequence flanking the other side of the insertion (see figure 2.14).

For these cross cloning site sequences, if we use BLAST searches or alignment tools like CLUSTAL the outcome will be utterly confusing, one would be misled to believe the insertion may occurred in two neighboring sites, at least the precision of the location will be lost. Our program maps and presents this situation clearly (Fig. 2.14).

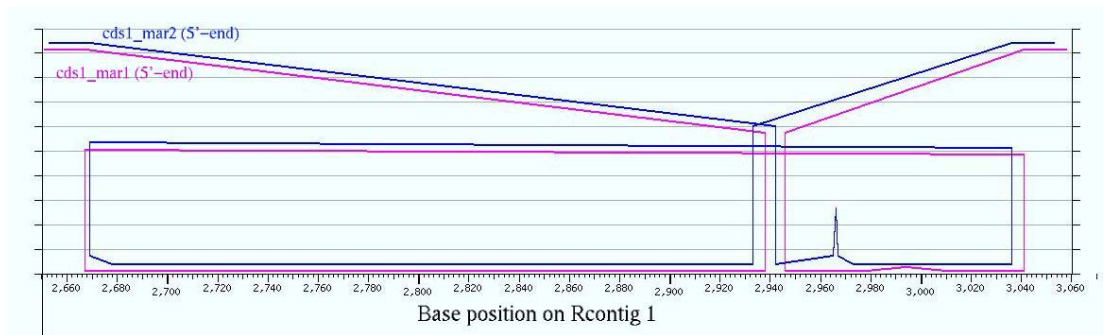


Figure 2.14. Each flanking sequence forms two stretches of homology with the contig1. cds1_mar1 (purple) matches the upstream stretch first, then the downstream one. The upstream homology is in the reverse direction, i.e. the complement of that stretch matches the contig1. The cds1_mar2 (blue) matches the downstream side first, then the upstream side.

Graphical Web Presentation

The insertion point-mapping program is intended to serve the biological community. Therefore accessibility is part of the design. At present, the most widely available method of access is the World Wide Web. For this reason the output of the presentation program is designed for web access. Once the Levenshtein distance matrices are established and filtered through the **L** limit and **H** limit, one can view the maps by visiting the web page at <https://129.15.160.110/DNAMapping/> (Note: the slash at the end of the line is required.)

The graph is generated on the fly when the user requests for an insertion map. The server keeps only the matrix of the homology. This approach saves storage space, especially when the insertion map database increases in size. The downside of this approach is every time a user requests an insertion map, the server has to regenerate the map. But user requests are distributed over time, it's unlikely for a surge of web requests to occur any time soon.

The graph-generating program is modeled from an open source program called PHPLOT by Afan Ottenheimer (It is now developed to version 5.0, freely available for everybody from SourceForge at <http://sourceforge.net/projects/phplot/>). It has been extensively modified to fit into this map-generating process. Several functions specific for map generation were added, and many other functions were modified. The modified PHPLOT takes in the matrix generated during the mapping process and calculates for positioning the map and scaling the X-axis.

The accepted matrix could be in one of two formats as in the following. Both formats track the base position for each stretch of homology and the level of homology.

```

array(
  array(key[0][0] => val[0][0], key[0][1] => val[0][1], key[0][2] => val[0][2], ...),
  array(key[1][0] => val[1][0], key[1][1] => val[1][1], key[1][2] => val[1][2], ...),
  array(key[2][0] => val[2][0], key[2][1] => val[2][1], key[2][2] => val[2][2], ...),
  ...
);

Or,

array(
  array ( src_length => #, contig_length => #, incr_frag => #, src#_homo# => homo_val,
    array( X => array( ***indexed array elements ***),
      Y => array( ***indexed array elements ***),
      label => array( ***Optional indexed array elements, each element could be an array ***),
      function => array( ***Optional indexed array elements, each element could be an array ***))
    ),
  array ( src_length => #, contig_length => #, incr_frag => #, src#_homo# => homo_val,
    array( X => array( ***indexed array elements ***),
      Y => array( ***indexed array elements ***),
      label => array( ***Optional indexed array elements, each element could be an array ***),
      function => array( ***Optional indexed array elements, each element could be an array ***))
    ),
  array ( src_length => #, contig_length => #, incr_frag => #, src#_homo# => homo_val,
    array( X => array( ***indexed array elements ***),
      Y => array( ***indexed array elements ***),
      label => array( ***Optional indexed array elements, each element could be an array ***),
      function => array( ***Optional indexed array elements, each element could be an array ***))
    )
  );

```

The graphical presentation program also staggers multiple sequences if they all map to the same contig or the same region of a genome (Figure 2.15). Each query sequence is also color coded. This makes the visual analysis clear and easy.

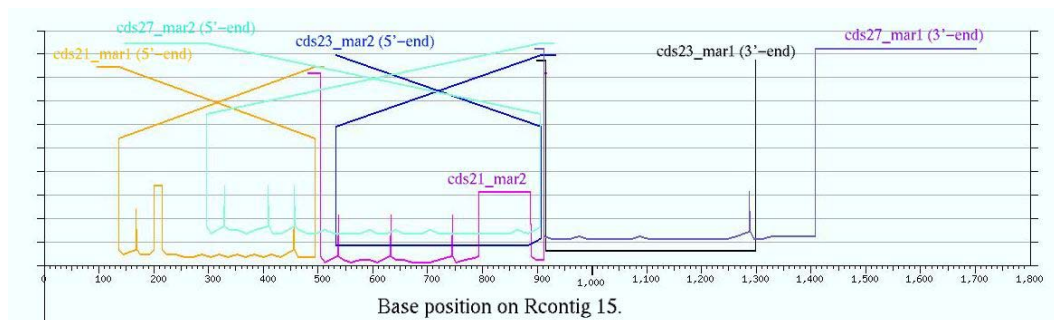


Figure 2.15. Layered presentation of the multiple query sequences matching a single contig. The sequence derived from primer mar2 of insertion cds21 (cds21_mar2) is the first layer. Its bottom is closest to the X axis. The sequence from primer mar1 of insertion cds21 (cds21_mar1) is the second layer. The bottom of cds11_mar1 is the second closest to the X axis. The cds23_mar1 is the third layer, the cds23_mar2 the forth, the cds27_mar1 the fifth, the cds27_mar2 the sixth.

More detailed use and interpretation (in biological terms) of the computer generated DNA insertion maps can be found in Chapter 1, Genome Wide Survey of Polysaccharide Biosynthesis Genes in *Myxococcus xanthus*. Some example output pictures follow.

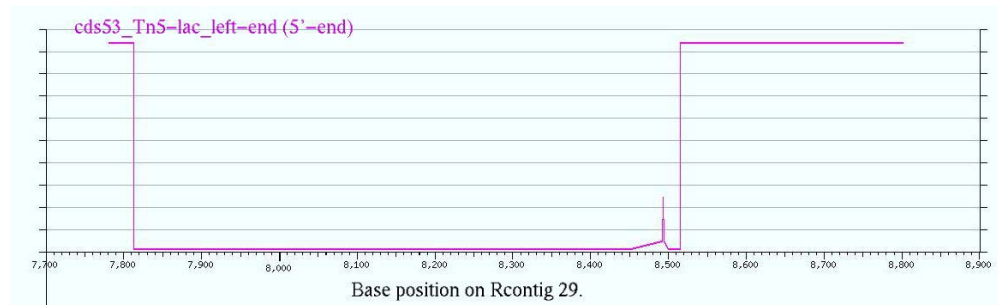


Figure 2.16 Only one side of the insertion is known. However, this single pass readout is of high quality for more than 630 bases. Since the sequence quality is high, the insertion point should be close to base 7810 position.

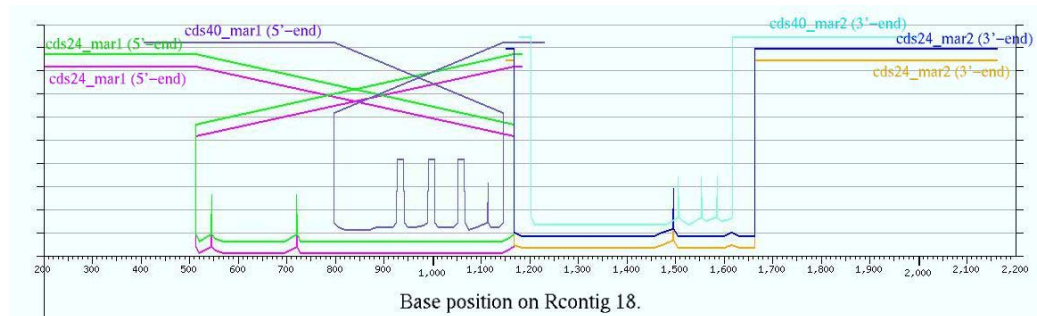


Figure 2.17 Duplicated insertion sequences in our database are automatically detected and clearly displayed. Shown here are two pairs of cds24 sequences, they are mapped to the same region of the same Rcontig.

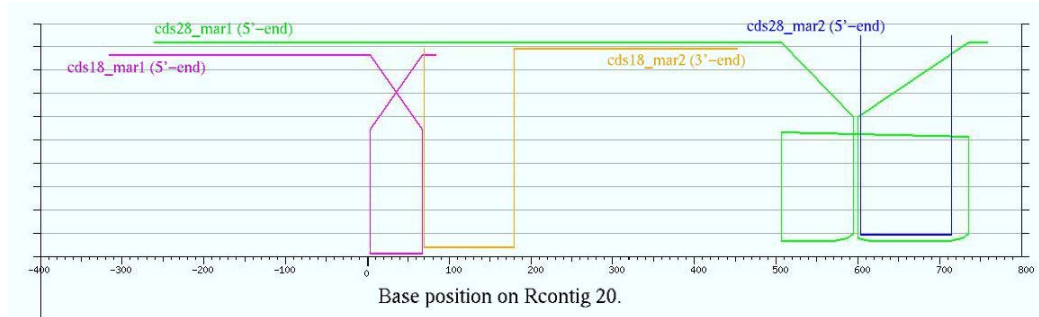


Figure 2.18 When Rcontig is limited on the 5' end, the cds18 mar1 and mar2 sequences become simple side by side arrangement.

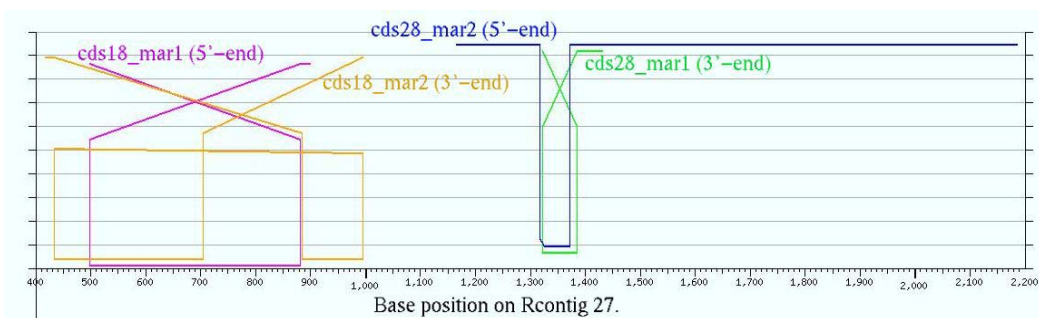


Figure 2.19 Same as in Figure 2.18, but the 5' end of the Rcontig is extended and allows the cds18_mar2 to match fully and display the far end that crossed the restriction site used in cloning digestion. Here the 3' end of the Rcontig is limiting, cds28 becomes un-solvably confusing.

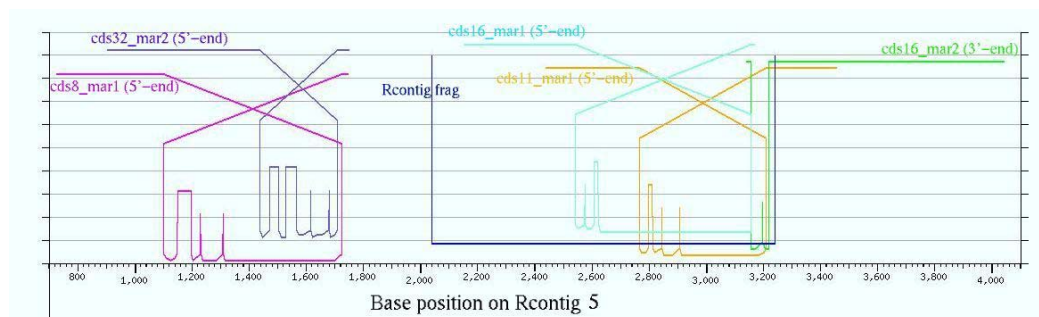


Figure 2.20 The Rcontig is limiting on the 3' end. The relative positions between cds16 and cds11 cannot be solved. After manual checking, cds32_mar2 homology here is an unreliable match. From the graph, it actually can be seen that the match never had a flat area at the bottom.

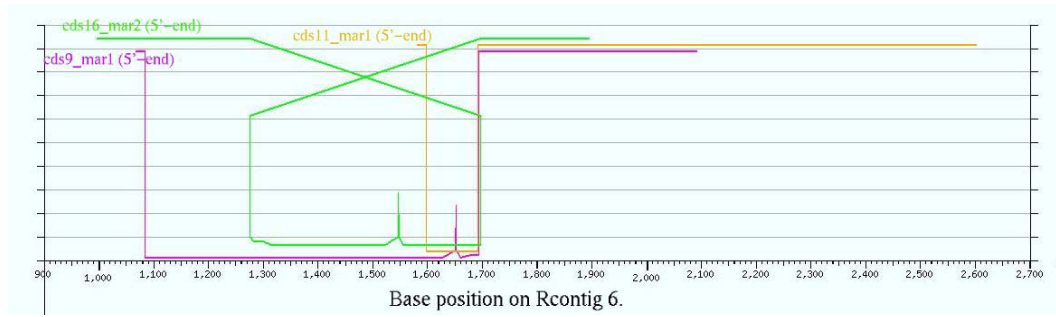


Figure 2.21 Similar to Figure 2.20, but focused on cds16 and cds11, they are long enough to bridge the gap with good confidence, and form a new Rcontig.

DISCUSSION

BLAST and CLUSTAL and their derivatives have been the mainstay for biological sequence analysis for more than ten years. They address the questions like what is the possible functions of a query sequence, or how a set of related sequences differ. However, there are questions in biomedical research that are not addressed by the presently available tools (computer programs). This chapter describes a new algorithm for sequence analysis attempting to solve issues about detailed sequence mapping, which are not addressed in BLAST and CLUSTAL and their derivatives (Kohli and Bachhawat, 2003; Chenna et al., 2003). This program produces a graphical presentation to symbolically display the similarity between the query sequence and the database sequence. The example used in this chapter as the basis for describing the problem is the insertion (DNA) point sequence mapping. However the sequence can be protein as well. It really depends on the issues at hand. I would like to call this sequence mapping tool SHAPE (Sequence Homology Annotation, Presentation and Editing). I intend to make this tool web user editable.

This program also solves a much broader sequence analysis issue that results from sequence shuffling due to various biological or artificial processes. For example, cloning of the transposon the way discussed in this article is a result of biologically disrupting a DNA sequence (transposition), artificially breaking and rejoining (shuffling, restriction digestion and ligation) of the sequence, and subjecting it to a biological process (transformation) to produce (amplify) the products (plasmid, including the insertion and flanking sequences). Breaking and rejoining sequences do happen as completely natural biological processes as well, such as messenger RNA splicing, intein (protein intron) splicing, transposition, and

genomic sequence inversions, *etc.* To be able to analyze and discover these processes with an automated tool like the one described here is valuable to biomedical researchers.

The algorithm for the SHAPE program is to 1) calculate the Levenshtein distance between the query sequence (including its complement) and the subject sequence within a sliding window, 2) find the pieces with the smallest distances (as potential homology) and integrate them into stretches of homology if they pass some artificial filter values. Then 3) output the information in a graphical form.

Although the algorithm used in this program is somewhat similar to the one used in BLAST, there are substantial differences: (1) There is only a symbolic alignment and base positions. The bases are not shown because it is not of primary concern. However, the base codes will be interactively available in a updated version of this mapping tool. (2) The gap or breakage in the alignment are symbolically connected therefore it needs no artificial gap penalties, nor artificial gap penalty weight matrix used in BLAST and CLUSTAL. And (3) since our SHAPE program uses the space much more efficiently it presents an integrated view of both global and local alignment and in a single picture. It is visually intuitive, more conducive to the user, while BLAST outputs a local alignment and usually spanning pages.

Compared to the BLAST program (Altschul et al., 1990, Wang and Mu, 2003) class of programs, this sequence mapping tool integrates all significant alignments into a contiguous graphical view. Users don't have to flip over pages of alignment to find out the start and end of each stretch of the alignments and mentally integrate the pieces of alignment into an overall homology. With the SHAPE program all pieces of alignment are presented in a

simple picture. It is visually direct.

Compared to the CLUSTAL class of programs (Altschul, et al., 1990; Kohli and Bachhawat, 2003; Chenna et al., 2003), the SHAPE program automatically aligns the complement of the sequence (or its fragments) if it has significant matches to the subject sequence. In addition, SHAPE also automatically align all query sequences matching a contig (Figures 2.15, 2.17, 2.18, 2.19, 2.20, 2.21). It presents the result in a single clear view. Therefore, SHAPE has the advantage of being uncluttered, direct and faster than CLUSTAL in processing and presenting results. The weakness of SHAPE is it does not have the ability to output matches to multiple contigs yet.

The sequence mapping tool has the following additional features that other tools do not have. First, it tolerates a lot of errors in the sequence. As presented in the RESULTS section of this chapter, to map an insertion site one needs only about one-hundred-bases contiguous homology or its equivalents. This capability made it possible to use error-laden sequences in insertion site mapping. For example, single pass sequencing yields several hundred bases of sequence with quite unpredictable quality. With the help of this mapping tool, single pass sequencing results are usually good enough to map the insertion sites (see Figures 2.15 and 2.17 for examples). For this project, 41 of the 52 attempted single pass sequencing jobs (for 26 templates, i.e. insertions) produced sequences. Among the 41, 38 were longer than one hundred bases (including undetermined base calls, designated as N). All 38 sequences were mapped to their correct locations (for 24 insertions).

Second, the base coordinates are automatically scaled and clearly labeled in a single view as

the result output. There is no need for scrolling the page. Third, the level of homology is represented by distinct symbols. The length of the homology, the position of the homology, the level of homology, and the direction of the homology are all presented in a simple picture. Finally, it layers out multiple query sequences if they match to the same contig or the same region of the genome (Figures 2.15, 2.17, 2.20).

Available programs use standard algorithms for sequence alignment. For example, in order to align just two sequences, it is standard practice to use dynamic programming (Needleman and Wunsch, 1970). This guarantees a mathematically optimal alignment, given a table of scores for matches and mismatches between all amino acids or nucleotides (e.g. the PAM250 matrix (Dayhoff et al., 1978) or BLOSUM62 matrix (Henikoff and Henikoff, 1992)) and penalties for insertions or deletions of different lengths. As the scope of sequence analysis expands and intensifies, more and more sequence analysis tool researchers pay increasing attention to flexibility and result presentation (Kohli and Bachhawat, 2003; Chenna et al., 2003). Coincident with this trend of modern sequence analysis tool development, our SHAPE integrates all the information into a single graphical presentation. Moreover, our SHAPE is going to be further enhanced with the ability of interactive editing, including interactively viewing and editing the base codes (and amino acid codes) and world-wide collaboration if a user so desires. (A preview of the enhancement is available by request.)

Another point that has never been addressed in other sequence analysis tools is made clearer in SHAPE to align both the forward and reverse strand of DNA sequences is necessary because the sequence readout from the sequencing primer might cross the cloning site

(Figures 2.12 and 2.13). In our real world mapping process, several such cases were discovered. The ability of the program to map forward and reverse direction of a sequence improves the precision of the map result (See Figure 2.14 for an example). It is expected that this sequence mapping algorithm could be used to potentially find and present graphically the domain rearrangements in many protein sequences, or other rearrangements that are not normally graphically discernable, such as small transposable elements and genomic inversions.

The deficiencies of the program includes: (1) it does not yet have the ability to dynamically use the many publicly available databases. (2) the implementation of the algorithm is only a prototype model, meaning it could be vastly improved by careful engineering to increase the speed and accuracy. (3) it does not yet have stable web address (needs a registered domain name to host the program.) (4) it has not yet been ported to a non-web based environment. That is, it cannot be run on a PC or a Mac. (5) it does not have any funding at all for its development and operation. And (6) due to the time constraint, the publicly available web access at present is limited to retrieving the insertion maps of the *M. xanthus* from Downard's lab. When time allows, it is going to be made possible for web users to enter their insertion sequences and genomic sequences, and generate their favorite insertion maps.

Future improvements to the SHAPE program will include (1) implement the algorithm in C or C++ language to improve computation speed. (2) Based on the improvement (1), this program can be made more portable. In other words, it is more likely to be run on a PC. (3) Reduce the density of the sliding windows. I have tried to slide the windows every two bases, and the speed improvement is almost proportional. (4) Make the filter value **L** and **H**

dynamically settable to increase sensitivity and reliability of the result. (5) Make it possible to track the exact bases that differ in the pair of sequences.

Future enhancement will be (1) make the output interactive for providing information and communication between users of the same sequence. A test prototype is available already. (2) make the sequence codes editable. (3) make the alignment automatically labeled, and interactively user modifiable.

APPENDIX

The system setup for running the SHAPE: MySQL from MySQL AB, Bangårdsgatan 8, S-753 20 Uppsala, Sweden, is used to make the databases. PHP from Apache Software Foundation (<http://www.apache.org/>) is used as the programming language. Apache also from Apache Software Foundation (<http://www.apache.org/>) is used as the web server. Linux is the Operating System for all developmental work. These programs are all GNU open source software. Thanks to all those involved in developing of these programs and making them available to all at no cost, and freely modifiable and redistributable. Without these developers there would be no SHAPE to talk about.

CHAPTER 3

A-SIGNAL TRANSDUCTION REQUIRES THE CATABOLIC PATHWAYS OF A-FACTORS

INTRODUCTION

It is now well accepted that bacterial cells don't live in isolation, they communicate constantly with each other and with their environment, just like the cells of higher organisms (Federle and Bassler, 2003). This cell-cell communication is based on signal mediators (molecules) that are exchanged between the cells. Usually bacterial signal molecules are dedicated species of chemicals produced specifically for the purpose of communication. Therefore, signal molecules have to be made, sent, then received, and responded to by the cells to complete a communication cycle. For example, N-acyl-homoserine lactone (also known as autoinducer 1, AI1) is a signal molecule made specifically for communicating / monitoring intraspecies population density in Gram-negative bacteria such as *Pseudomonas* (Eberl, 1999); while furanosyl borate diester (also known as autoinducer 2, AI2) is for monitoring the interspecies population density (Federle and Bassler, 2003). Communication for this purpose is called quorum sensing and appears to be widespread. Due to technical limitations, we are able to study the signaling systems employing large amounts of signal molecules (micromolar level) only such as quorum-sensing, although there is evidence that other vital signals function at lower concentrations. For example, the resuscitation-promotion-factor (Rpf) functions at picomolar level in *Micrococcus luteus* and many *Mycobacterium* species (Mukamolova *et al.*, 1998; Shleevea *et al.*, 2003, 2004; Zhu *et al.*, 2003; Tufariello *et al.*, 2004). The signals in cell-cell

communication are almost always involved in viability, pathogenicity, and resistance to the harsh environment.

M. xanthus is a nonpathogenic model organism well suited for studying the cell-cell communication systems. A wealth of knowledge has been accumulated in the understanding of signaling processes in *Myxococcus xanthus*. Once densely packed cells are subjected to starvation on a stable surface, well-coordinated cell-cell interaction takes place and culminates in the construction of a well defined spheroid multicellular structure called a fruiting body (Shimkets and Kaiser, 1982; Shimkets, 1990; Kaiser and Losick, 1993; Downard *et al.*, 1993; Dworkin, 1996). After cells aggregate into fruiting bodies, individual rod-shaped cells within these structures begin to differentiate into spheroid-shaped spores that are resistant to environmental stresses such as heat, desiccation, sonication, *etc.* Thus, the *M. xanthus* development cycle occurs in a series of steps that include starvation, aggregation, fruiting body formation and sporulation.

Multicellular development in *M. xanthus* is coordinated by cell-cell communication, and five putative signaling systems have been identified based on analysis of mutants deficient in cell-cell communication (Hagen *et al.*, 1978; LaRossa *et al.*, 1983; Downard *et al.*, 1993). Like five channels of signals delivered to an orchestra, they have to be integrated inside the cell to proceed with the development program. Study of *M. xanthus* cell-cell communication has been facilitated by the isolation of signaling mutants that are unable to complete development by themselves, but they can overcome their developmental defects if they are mixed with wild-type cells. This extracellular complementation does not involve a permanent genetic exchange from wild-type cells to mutant cells, as the mixed culture

fruiting bodies contain spores that retain their respective phenotypes (wildtype or mutant) in subsequent analysis. The hypothesis is that these *M. xanthus* mutants are defective for signal production but can respond to the signals produced by the wild-type cells. They can be classified into five classes: *asg*, *bsg*, *csg*, *dsg*, and *esg*. Mutants from the same class fail to complement each other, whereas mutants from different classes can reciprocally rescue their partner to complete their co-development process.

Several *asg* mutants that fail to produce the A-signal have been well characterized. The defects observed in the *asg* mutants result from lesions in one of five genes, *asgA*, *asgB*, *asgC* (Kuspa and Kaiser, 1989), *asgD* (Cho and Zusman, 1999), and *asgE* (Garza et al., 2000). The *asg* mutants fail to produce and release normal levels of A-factor, or A-signal, which is composed of seven amino acids (tyrosine, proline, phenylalanine, tryptophan, leucine, isoleucine, and valine) as its major constituents, and nine other amino acids (Kuspa et al., 1992a; Plamann et al., 1992). The proteolysis target is not identified, but it must be some protein(s) on the cell surface or those continuously released into the medium because there is no other protein available in the standard development medium. The observed facts are, the more concentrated the cells are in the development medium, the more concentrated the proteases; therefore the higher the concentration of the amino acids released into the development medium, and the higher the A-factor concentration (Kuspa and Kaiser, 1992b; Kim and Kaiser, 1992). Since the signal producer and receiver are the same set of cells, this kind of signal is said to be an autocrine (paracrine) signal, typical of most bacterial signals. The overall net effect is that the A-signal strength is directly proportional to the protease concentration and to the cell concentration (Kim and Kaiser, 1992). Because A-signal strength (A-factor activity concentration) is cell density dependent, A-signal sensing is

called quorum sensing, characteristic of the autocrine mode signals. Consequently, A-signal is known as a quorum-sensing signal. A-factor activity assay is experimentally measured by monitoring the A-factor dependent promoter Ω 4521 using a transcriptional β -galactosidase fusion in an A-signal defective background.

Quorum sensing is necessary for proper development in *Myxococcus xanthus* because without a high enough concentration of cells in the immediate environment, it would be difficult to recruit enough cells into the aggregation centers to form fruiting bodies. If fruiting bodies are not formed properly, the sporulation process will be inhibited. The minimum effective A-factor amino acid concentration is 10 μ M (individual amino acids or a mixture of the seven major A-factor amino acids, see below) (Kuspa *et al.*, 1992b). Consequently, the purpose of the development would be defeated. However, if the A-factor concentration is too high (above the maximum concentration that is required for maintaining stringent response control), the cells respond with growth, instead of development, even when cell concentration is as high as in colonies on agar plates.

Interestingly, not all amino acids are A-signal mediators, and the A-factor activity is cumulative. That is, the total of the A-factor activity is the sum of the A-factor activity contributed by each A-signal amino acids in the medium. However, different amino acids are present at very different concentrations in the development environment. The ones that are at very low concentration, such as glutamate, obviously contribute little to the A-factor activity. Some amino acids, such as cysteine, even have negative effect. Therefore, the major contributor amino acids are limited to seven (tyrosine, proline, phenylalanine,

tryptophan, leucine, isoleucine, and valine). The seven major A-factor amino acids have higher specific A-factor activity (inducing more A-factor activity per amino acid molecule). Nevertheless, amino acids do not account for the total A-factor activity in the developing culture. Yet A-signal study has been primarily focused on the A-signal production and releasing aspects. Little is known about how the cells respond to A-signal. Consequently, this becomes the focus of research for this chapter.

A simple analysis of the seven most active A-factor amino acids reveals that A-signal amino acids are chemically heterogeneous in nature. Leucine, isoleucine, and valine are hydrophobic and aliphatic with small side-chains. Tryptophan, tyrosine and phenylalanine are also hydrophobic, but aromatic with bulky side chains. Proline's structural backbone is strained with a looped-back side chain. Simply, there is no common feature for these seven amino acids that might allow interaction with a cell surface or intracellular receptor to activate the A-signal response process. Since the scarcity of amino acids for protein synthesis produces the stringent response and initiate the development process, we reasoned (more in the discussion section), the most likely candidate for A-signal processing involving amino acids is the amino acid degradation pathway. In addition, *M. xanthus* normally grows on amino acids, obviously capable of catalysis of amino acids. Besides, because we have the branched-chain α -keto acid dehydrogenase (BCKAD, an enzyme required for BCAA degradation) mutant *esg* at hand (Downard *et al.*, 1993), we tried to test whether the A-signal processing actually uses this BCAA degradation pathway. If it does, the *esg* mutant should produce no A-factor activity in response to the BCAAs while the wildtype should produce normal amount A-factor activity in the same assay. The results supported our

hypothesis. A number of other A-signal amino acids seem to need their own degradation pathways for generating A-signal response, as measured by A-factor activity assay. Then we tried chemicals other than amino acids, and found short chain fatty acids and pyruvate also have A-factor activity.

In this Chapter, I will present evidence that suggests for the first time that the A-signal response process in the *M. xanthus* signal transduction system requires the catabolic breakdown of the A-factor amino acids using the same system as used for growth on these amino acids. Based on these findings, a model is proposed to integrate my new findings with what have been accumulated in the literature about the A-signaling process.

MATERIALS AND METHODS

Strains Used

The wildtype *M. xanthus* strain (DK6600) for A-signal activity assay carries the structure $\Omega 4521$, which is a *lacZ* gene fusion to an A-signal dependent promoter, and *asgB480*, which eliminates A-signal production, and makes the strain DK6600 depend on exogenous A-signal for A-signal response and development. The developmentally regulated gene blocked by $\Omega 4521$ codes for a serpin (serine-protease), and is dispensible for development. This strain is used as a standard assay setup for A-signal activity (Kuspa *et al.*, 1992a). All strains where A-signal response is measured carry the $\Omega 4521$ and *asgB480* in addition to the gene mutation under investigation. This includes the wildtype DK6600, *esg*, *pah*, *aldA*, and *dcm-1*. Strain *dcm-1* carries a TnV insertion mutation in the propionyl-CoA carboxylase gene (Yoshio *et al.*, 1997), which is required for isoleucine catabolism. Strain *esg* carries a Tn5 insertion in the E1 α gene for branched chain α -keto acid dehydrogenase (Downard *et al.*, 1993). The strain *pah* carries an interruption in the phenylalanine hydroxylase gene, which converts phenylalanine to tyrosine, and is required for the catabolism of phenylalanine. The *aldA* strain is created in Zusman's laboratory (Ward *et al.*, 2000). The gene *aldA* codes for the enzyme alanine dehydrogenase, which converts alanine to pyruvate, and is required for the alanine catabolism.

Internal Fragment Replacement Mutagenesis

The protocol is essentially the same as described in the Chapter 1 except using different primer pairs for different target genes. An internal gene fragment was PCR amplified, and

cloned into a vector, then electroporated into a *M. xanthus* wildtype DK6600. The primer pairs used and the genes targeted are listed in the table below (Table 3.1).

Table 3.1 Primers used for internal fragment replacement PCR mutagenesis

Target gene	Primer Sequence	Result frag. (bp)
<i>pah</i>	For 5' -CG GAATTC GCGGACCAGGCCG-3'	1108
	Rev 5' -CG GAATTC CAGCGACAGCTTGA-3'	
<i>aldA</i>	For 5' -CGGACGAGGTCTGGAAGCGC-3'	750
	Rev 5' -AGGTGGACGTCTGCGGCACG-3'	

A-Factor Assay

Production of β -galactosidase from the structure Ω 4521 (a Tn5*lac* insertion) is controlled by a developmentally regulated promoter that requires a functional A-signaling system (Kuspa and Kaiser, 1989), or exogenous A-signal. When Ω 4521 is present in an *asg* mutant background (such as *asgB480*, which produces a very low A-signal level), the β -galactosidase level is very low during development unless wild type *M. xanthus* cells (which release the A-signal) are mixed in or a substance that has A-factor activity is added. This phenomenon is conveniently adapted for quantification of A-factor. Therefore, A-factor activity is calculated from the production of β -galactosidase from Ω 4521-*lac* fusion in the test strain.

Exponentially growing test cells were sedimented and washed by resuspending the cells in approximately one volume of MC7 buffer (10 mM morpholinepropanesulfonic acid, 1 mM

CaCl₂, pH 7.0) at room temperature. After 10 min at room temperature, the cells were sedimented and resuspended in MC7 to a calculated density of 5×10^9 cells per ml. Aliquots of 25 μ l (1.25×10^8 cells) were added to wells of a 24-well microtiter plate; each well contained 400 μ l of MC7 buffer or MC7 buffer plus a substance to be tested for A-factor activity, bringing the total volume to 425 μ l. These plates were incubated for 20 h at 32°C in a humid chamber before the assay for β -galactosidase activity. (Under *asg*⁺ background, the β -galactosidase from Ω 4521 reaches the maximum at approximately 20 h.) One unit of A-factor activity is defined as the amount required to stimulate the test cells to produce 1 Miller Unit of β -galactosidase activity (1 nmol of o-nitrophenol per min) above background. A-factor activity was calculated from the linear region of a dose-response curve for each fraction tested.

Other Conditions

Growth conditions and development conditions are the same as have been described in Chapter 1.

RESULTS

Branched chain keto acid dehydrogenase is required for *M. xanthus* response to the three branched chain amino acids

M. xanthus has been shown to use a number of amino acids as a quorum-sensing signal, the A-signal. How cells respond to this chemically heterogeneous signal remains unknown although it is clear that the ultimate target of the signaling system is the activation of the transcription of a group of A-signal-dependent genes. The gene containing the $\Omega 4521$ Tn5lac insertion, that places β -galactosidase production under control of the A-factor dependent promoter of that gene, is an example that these genes activated in response to high cell density. Since *M. xanthus* is capable of growth using amino acids as the primary carbon and energy source, the hypothesis that amino acid catabolism is involved in the response to the A-factor amino acids was considered. Two of the branched-chain amino acids (BCAA), leucine and isoleucine, were previously shown to be important components of A-factor. To test the requirement for BCAA catabolism in the A-factor response, an *esg* mutant strain was utilized. This mutant has been shown to be defective in the production of branched-chain keto acid dehydrogenase (BCKAD) that catalyzes an early step in the degradation of all three of the BCAA (valine as well as leucine and isoleucine). As shown in Fig. 3.1, while wild-type cells were able to activate $\Omega 4521$ expression in response to isoleucine, the *esg* mutant failed to respond to this amino acid. The defect in the *esg* mutant was specific in this strain retained a normal level of A-factor response to another important A-factor amino acid, phenylalanine.

Individual branched chain amino acids were added to the wildtype and the *esg* mutant assay buffer for A-factor activity, respectively. The *esg* strain carries a Tn5 insertion mutation in

the gene branched chain α -keto acid dehydrogenase (BCKAD). After 20 hours of incubation, the β -galactosidase activity is measured. The *esg* mutant was completely unable to generate A-signal response to isoleucine while the wildtype responded to isoleucine normally (Figure 3.1). In the meantime, both the wildtype and the BCKAD deficient *esg* mutant produced identical, roughly linear response in A-factor activity to the full range of tested concentration of phenylalanine from 0 to 1.25 mM. This result demonstrates that BCKAD is required for producing the A-factor activity in response to the BCAA isoleucine. Since phenylalanine is not a BCAA, its degradation does not need the enzyme BCKAD, therefore *esg* produced the same amount of A-factor activity as the wildtype in response to phenylalanine (Figures 3.1 and 3.2).

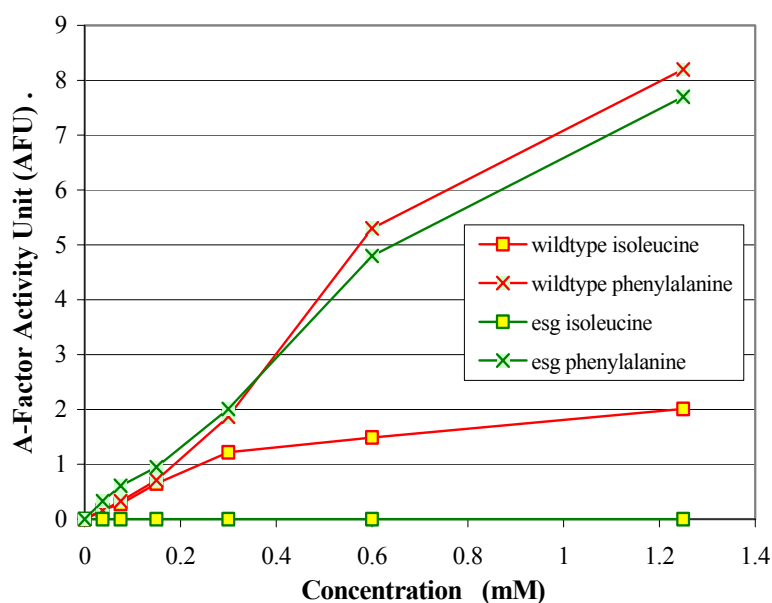


Figure 3.1 Comparison of A-factor activities between the wildtype and the *esg* mutant in response to branched chain amino acid (BCAA) leucine. Both wildtype and *esg* mutant responded identically to phenylalanine, whereas only the wildtype is able to respond to isoleucine. *esg* mutant is completely unable to respond to isoleucine.

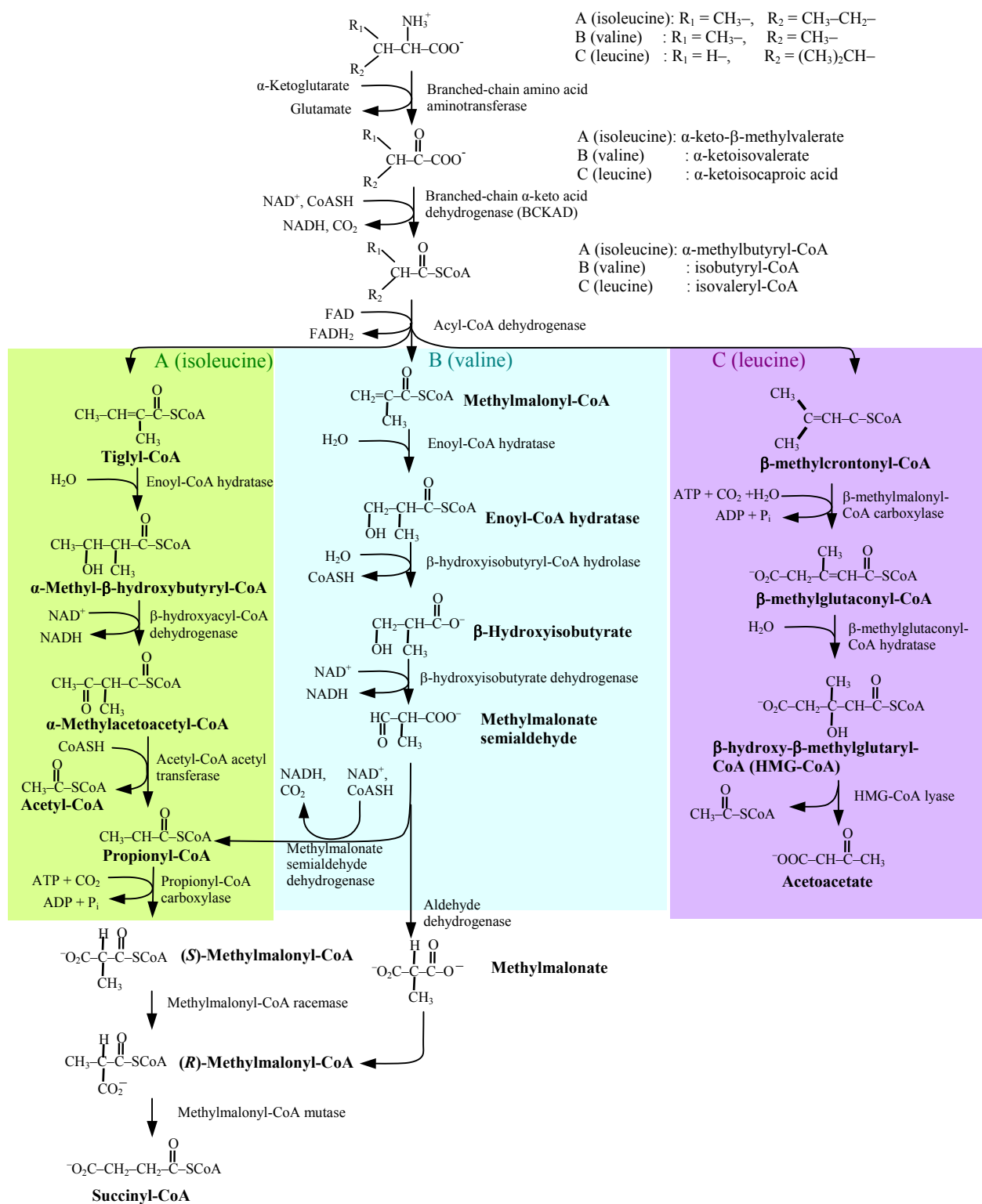


Figure 3.2 Degradation pathways for the branched amino acids. The end products (succinyl-CoA, acetyl-CoA, acetoacetyl-CoA, and acetoacetate) are fed into the central metabolism system: TCA and glyoxylate cycles.

The *esg* mutant also failed to respond to the two other BCAAs leucine and valine (Figure 3.3). It is worth noting though that although the AFU value is relatively small, the amount of A-factor activity elicited with valine is considerably greater than previously reported (Kuspa et al., 1992b). Therefore, this makes all three branched chain amino acids potent A-signal molecules (more results below).

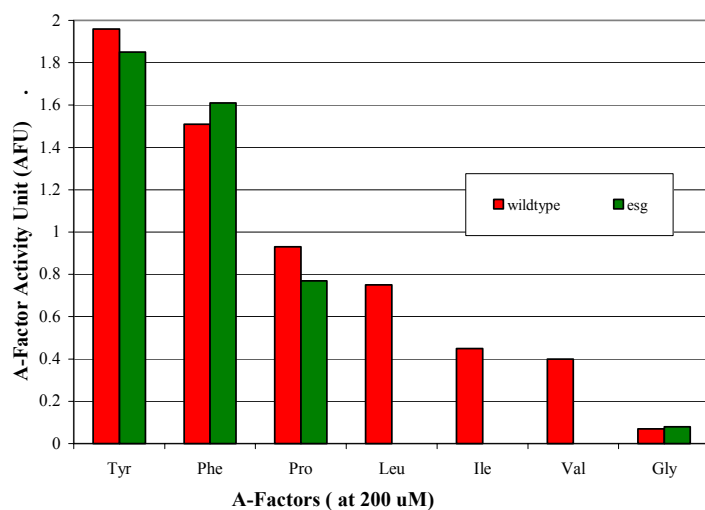


Figure 3.3 Comparison of A-factor activities of selected amino acids between the wildtype and the *esg* in response to individual amino acids, each at 200 μ M.

The activities observed here generally are comparable with the previously published (Kuspa *et al.*, 1992a) data on A-factor amino acids, except that valine shows a similar level of activities to leucine and isoleucine much higher than previously reported undetectable level (Figure 3.3). Figure 3.3 also shows that the same amino acids at the same concentration as in wildtype cells, *esg* mutant did not produce A-signal activity in response to three branched chain amino acids: leucine, isoleucine, and valine. This is as predicted in the Figure 3.2 since the *esg* mutation is in the branched chain keto acids dehydrogenase gene which is required for catabolism for all three branched chain amino acids. The deficiency in branched chain amino acids degradation in the *esg* strain causes severe defects in A-signal

response that cannot be overcome at any concentration of the BCAA isoleucine (leucine and valine behaved similarly, data not shown), while other amino acids, e.g. phenylalanine, function normally.

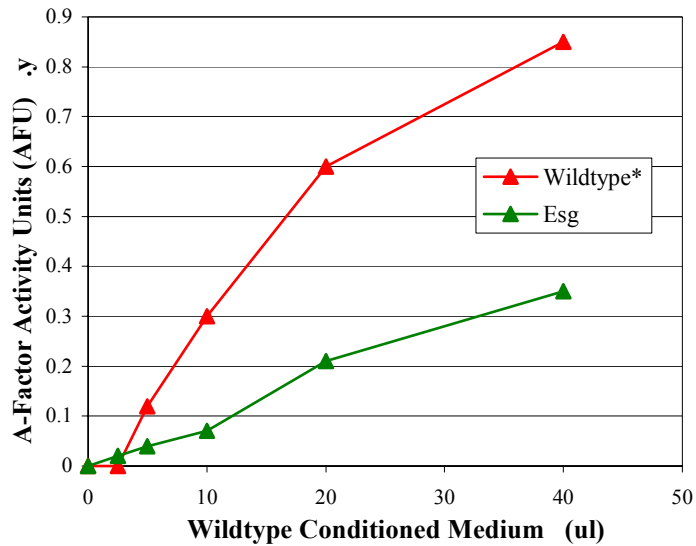


Figure 3.4 Wildtype cells (5×10^9 cells per ml) were incubated in MC7 (morpholinepropanesulfonic acid [pH 7.0], CaCl_2) for 2 to 4 hours with shaking. Then the cells were removed by centrifugation, the conditioned medium (supernatant) was added to the wildtype and *esg* in liquid development medium, and the A-factor activity for each strain was measured.

Since *esg* is defective in response to the branched chain amino acids as A-factor, we want to know find out how much A-factor activity is lost compared to the wild-type using the natural A-factor, which consists of most amino acids and some unknown compounds. To quantitate the amount of A-signal the *esg* strain produces under normal development conditions, wild-type cells (DK1622) (5×10^9 cells per ml) were incubated in MC7 (morpholinepropanesulfonic acid [pH 7.0], CaCl_2) for 2 to 4 hours with shaking. Then the cells were removed by centrifugation, the conditioned medium (supernatant) was added to A-factor activity assays for DK6600 (wildtype for A-factor activity assay) and the *esg*,

respectively. After 20 hours of incubation, the β -galactosidase activities were measured. This assay showed that the *esg* strain produced about half of the wildtype level A-factor activity in response to the wildtype conditioned medium (Figure 3.4), which is many times more than the minimum required to initiate the wildtype development (Kuspa *et al.*, 1992b).

Loss of propionyl-CoA carboxylase eliminates A-signal response to isoleucine and short chain fatty acid valerate

The results presented above are consistent with a model in which amino acid catabolism is required for the response to the A-factor amino acids. To further test this hypothesis, a mutant with a defect in another BCAA catabolism gene, propionyl-CoA carboxylase, was tested for the response to BCAA. The available strain for this test is the *dcm-1* strain described by Kimura and colleagues (Kimura *et al.*, 1997). The strain *dcm-1* carries a transposon insertion in the gene propionyl-CoA carboxylase (*pccB*), which has been shown to abolish the propionyl-CoA carboxylase activity. Propionyl-CoA carboxylase is involved in the step where the isoleucine degradation intermediate propionyl-CoA has to be converted to methylmalonyl-CoA in order to carry through the degradation process to succinyl-CoA and feed into the tricarboxylic acid (TCA) cycle (Figure 3.2). In some bacteria, this enzyme is involved in isoleucine and valine catabolism. In others, it is required only for isoleucine catabolism.

Three different concentrations of the branched chain amino acids leucine, isoleucine and valine were used as A-factor. The A-factor activity is reported as the percentage of the A-factor activity for the wildtype strain in response to respective BCAA. The result (Table 3.3) shows that the *dcm-1* strain was most defective in response to isoleucine in the A-factor

activity assay and activity was detectable only at the highest concentration tested. However, the insertional mutation in the gene *pccB* is known to cause multiple other defects in the strain *dcm-1*. Previous experiments showed that propionyl-CoA carboxylase is required for development, and *pccB* defect led to diminished sporulation efficiency, in addition to lowered levels of long chain fatty acids (C₁₆ to C₁₈) production during development (Kimura *et al.*, 1997). The *dcm-1* mutant also had more minor deficiencies in response to the BCAA, leucine and valine. It is likely that this mutant exhibit pleiotropic defects relating to the *pccB* mutation. These more minor defect in response may not be significant.

Table 3.3 Relative A-factor activity of the *dcm-1* strain compared to the wildtype DK1622.

Concentration (μ M)	Leucine (WT%)	Isoleucine (WT%)	Valine (WT%)
100	40	ND	73
200	54	ND	63
400	52	9	51

Note: Compare A-factor activities from the *dcm-1* strain with the wildtype in response to BCAA. The *dcm-1* strain is not responsive to the stimulation by isoleucine, whereas largely active in response to leucine and valine. ND: not detectable.

It is worth pointing out that in some organisms, especially higher organisms, valine degradation may take the route via propionyl-CoA intermediate. We searched the *M. xanthus* genome database (albeit still incomplete at this time), and found that there is a methylmalonate semialdehyde dehydrogenase homologue similar to the one from *Bdellovibrio bacteriovorus* with an E value at e^{-150} , at 1.30 Mbp on the *M. xanthus* genome physical map (Figure 1.10), which if proved to be true would convert methylmalonate semialdehyde to propionyl-CoA, and make the enzyme propionyl-CoA carboxylase necessary for completing the valine degradation pathway. However, with the same method I also found that there is an aldehyde dehydrogenase homologue similar to the one from

Streptomyces coelicolor A3(2) and many other species with the E value at 0.0, at 6.60 Mbp on the physical map (Figure 1.10), which if proved to be true would allow the valine degradation to bypass the requirement of the enzyme propionyl-CoA carboxylase, by converting the degradation intermediate methylmalonate semialdehyde to methylmalonate, methylmalonyl-CoA, then succinyl-CoA, feeding the TCA cycle. If both of those two enzymes are proven to be true to their predicted functions, they present a choice to the *M. xanthus* cell when and how to utilize the two branches of the final stage of valine degradation. The controlling condition for this choice could be development or vegetative state of the cell, nutrient richness, growth rate, or many other possibilities. From our A-signaling activity assay results, we think that the choice for the development state is bypassing the propionyl-CoA step. However the valine induced A-signal response is at ~70% of the wildtype level, therefore there is a chance that a fraction of the valine degradation intermediate may actually pass through the propionyl-CoA branch.

Phenylalanine hydroxylase is required for the A-factor response to phenylalanine

The first step in phenylalanine catabolism is catalyzed by phenylalanine hydroxylase. In *M. xanthus* a putative phenylalanine hydroxylase gene was previously identified based on DNA sequence analysis (Sun and Shi, 2001). A mutant lacking the phenylalanine hydroxylase gene *pah* was produced for this study and the mutant was tested for the A-factor response to phenylalanine. The *pah* mutant was constructed using the targeted PCR Mutagenesis method as described in the Materials and Methods section. The *pah* mutant carries two truncated copies of phenylalanine hydroxylase gene with the vector pZerO-2 inserted in between. Phenylalanine hydroxylase converts phenylalanine to tyrosine in the phenylalanine degradation pathway. Results presented in Figure 3.5 demonstrates that the *pah* mutant

showed a dramatically reduced A-factor response to phenylalanine, presumably due to its deficiency in degradation of phenylalanine. Also note that the basal level does not change as the concentration of the phenylalanine increases. In addition, the *pah* mutant responded to tyrosine normally (data not shown), indicating that the *pah* defect is specific for the A-factor activity in response to phenylalanine. Since previous sequence analysis (Sun and Shi, 2001) indicated that *pah* is the last gene in the operon, the A-signal response defect reported here is most likely due to the loss of the functional phenylalanine hydroxylase.

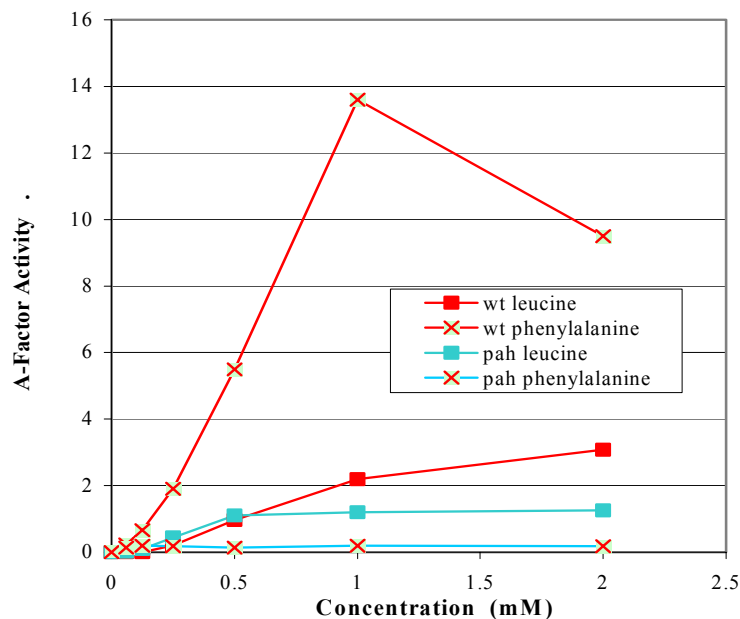


Figure 3.5 Compare A-factor activities from the *pah* strain with the wildtype in response to leucine and phenylalanine.

Alanine dehydrogenase mutation eliminates A-signal response to alanine, and pyruvate has A-factor activity

The *M. xanthus aldA* gene codes for the enzyme alanine dehydrogenase (Ward *et al.*, 2000), which converts alanine to pyruvate as the first step in alanine catabolism. The *aldA* strain was constructed with the targeted PCR mutagenesis method by Ward and colleagues,

resulting in a kanamycin resistant *aldA* strain (Ward *et al.*, 2000). If A-signal response does indeed require the catabolism of A-signal molecules, we reasoned, inactivating the enzyme alanine dehydrogenase to block the catabolism of alanine would block the A-signal response to alanine. Our results confirmed this. The *aldA* strain has completely lost the A-signal response to alanine in A-factor activity assay. As a surprise, we tested pyruvate for A-factor activity, and found pyruvate indeed generates significant amount of A-factor activity, in both the wildtype DK6600 and the *aldA* strain (Figure 3.6). This is a significant discovery because first, previously it was reported that pyruvate does not have A-factor activity. Second, pyruvate is not an amino acid, which implies that non-amino acid compounds can have A-factor activity. Third, pyruvate is an intermediate for many compounds' metabolic processes, such as amino acids (degradation product of all gluconeogenic amino acids, i.e. all 20 regular amino acids except leucine and lysine, and synthesis precursor for several, including alanine) and lactate (which could be reversibly converted to pyruvate by lactate dehydrogenase although unknown in *M. xanthus*). Although the glycolysis pathway does not seem to be functional in *M. xanthus*, the gluconeogenesis is known to be active, especially under starvation-induced development conditions, synthesizing huge amount of trehalose (McBride and Zusman, 1989). Therefore, sufficient amount of pyruvate must be available in the cell. As far as starvation-induced development condition is concerned, all ketogenic compounds can be glucogenic with the help of the glyoxylate cycle. Thus fatty acids catabolic products, acetyl-CoA, can be converted via glyoxylate cycle enzymes into succinate, which is further converted to fumarate, malate and pyruvate by TCA cycle enzymes. (See the Discussion section for more on this topic.) Finally, pyruvate is one of the major carbon and energy source in the chemically defined A1 medium for *M. xanthus*. However, *M. xanthus* grows on A1, instead of development.

Pyruvate under this condition (in A1 medium) does not function as the A-factor. (See the Discussion section for more details.)

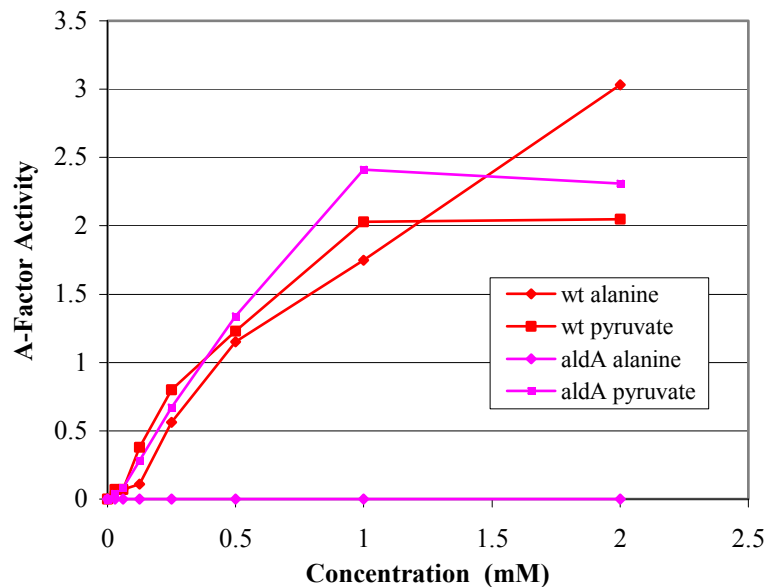


Figure 3.6 A-factor activities from the *aldA* strain compared with the wildtype in response to alanine and pyruvate.

Compared with the wildtype, the *aldA* strain completely lost A-signal response to alanine, while retained its response to pyruvate (Figure 3.6). It should be noted that defects caused by interruption in the gene *aldA* are multifold. Besides the loss of A-signal response to alanine described in this work, it has been reported (Ward *et al.*, 2000) that the *aldA* mutant does not develop normally under the standard starvation condition for development. A-signal response defect since alanine is not a major component of A-factor, and the *aldA* mutant can presumably respond to the other A-factor amino acids.

Arginine and two of the urea cycle degradation intermediates have A-factor activity

Several pathways are used for arginine degradation in bacteria (Morris, 2004; for a recent

review) and several intermediates in those pathways are commercially available. Since there is no information available about *M. xanthus* arginine catabolism, I decided to test four of those compounds for A-factor activity. The results in the Figure 3.7 show that both ornithine and citruline have good amount of A-signal activity, while putrescine and agmatine do not have any A-signal activity at all.

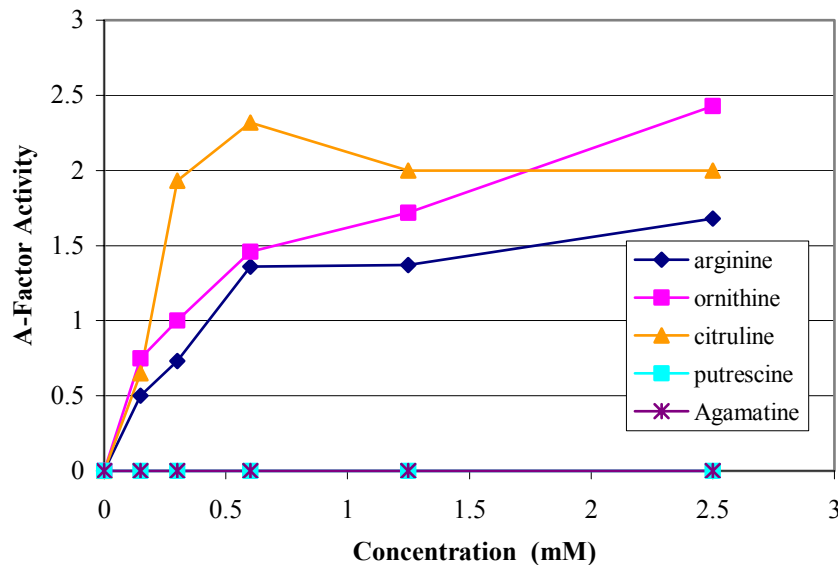


Figure 3.7 A survey of A-factor activity of arginine and its urea cycle degradation intermediates.

The fact that multiple arginine degradation intermediates have A-factor activity suggests that *M. xanthus* possesses a functional urea cycle. Consistent with this assumption was the discovery that a search in the *M. xanthus* genome found the urea cycle enzymes *argF* (ornithine carbamoyl transferase), *argG* (argininosuccinate synthase), and *argH* (argininosuccinate lyase) present in an apparent operon *argFBEHG_C^l* (the underline represents an unknown gene) at the 6.7 Mbp position according to the *M. xanthus* genome

¹ Interestingly, Harris and Singer (Harris and Singer, 1998) reported that there is an *argE* gene in the *asgE* region (see Figure 3.11). This is different from the *argE* discussed earlier, therefore I name it as *argE^{lone}* to distinguish it from the one in the *arg* operon. Their experiments show that null *argE^{lone}* would make *M. xanthus* an arginine auxotroph, and develop on A1 agar (which lacks arginine in the recipe). The relationship between the two versions of *argE* is not clear.

map (He *et al.*, 1994) (Figure 1.10). The other urea cycle enzyme (arginase) is found in the genome database at 5.79 Mbp according to the *M. xanthus* physical map. Nevertheless, arginine degradation is likely via arginase, *argD* (ornithine aminotransferase, at 0.10 Mbp on the physical map) and Δ^1 -pyrroline-5-carboxylate dehydrogenase (at 7.60 Mbp) to proline (Figure 3.8). Proline is known as one of the six major A-factor amino acids; presumably it is also degraded and fed into the TCA cycle. This shows that the urea cycle intermediates can be degraded into the TCA cycle intermediate fumarate or into proline, which in turn is degraded and fed into the TCA cycle.

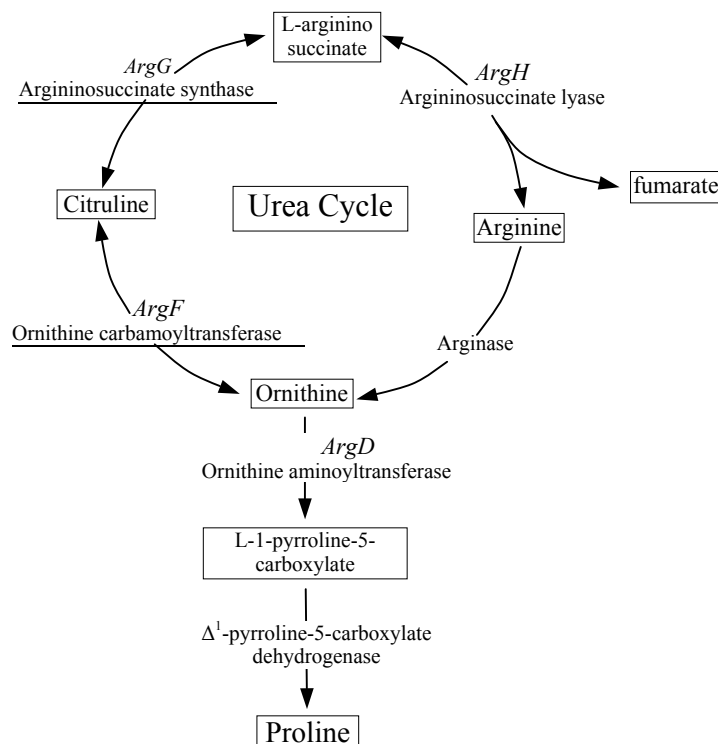


Figure 3.8 The urea cycle and arginine degradation pathway in *Myxococcus xanthus*.

Short chain fatty acids have A-factor activity

Branched chain amino acid degradation shares part of the pathway with the short chain fatty acid degradation. In reality, BCAA are converted to short chain fatty acids in the degradation pathway (Figure 3.2). The enzyme propionyl-CoA carboxylase is involved in short chain fatty acid degradation as well. When the short chain fatty acids butyrate (even-chain, C₄) and valerate (odd chain, C₅) are added to the A-factor assay medium with wildtype cells, a dose responsive curve shows the A-factor activity spanning from 60 μ M to 1 mM. In the *dcm-1* strain, butyrate induced a response similar to the wildtype, but valerate induced no A-signal response at all. The strain *dcm-1* lacks propionyl-CoA carboxylase, which is required for odd-chain fatty acid (valerate) degradation, but not needed for even-chain fatty acid (butyrate) degradation. Therefore, the *dcm-1* strain remains responsive to even-chain fatty acid butyrate, but not to the odd chain fatty acid valerate (Figure 3.9). However, it is not clear whether valerate and butyrate are part of the naturally produced A-factor mix. Consequently, it is not clear whether these two short chain fatty acids are part of A-signal, although they have good potential to be one.

At this time, it is still not known whether fatty acids with longer chains elicit the A-factor like response. But these results further indicate that A-factor composition might be more complex than previously indicated in that compounds beside amino acids that can be used as carbon and energy sources may be present and contribute to the total A-factor produced by cells. Presently, there is no evidence that any such compounds are produced at high concentration. In addition to amino acids, fatty acids and pyruvate (discussed earlier together with alanine) have now become new members the “A-factor club”.

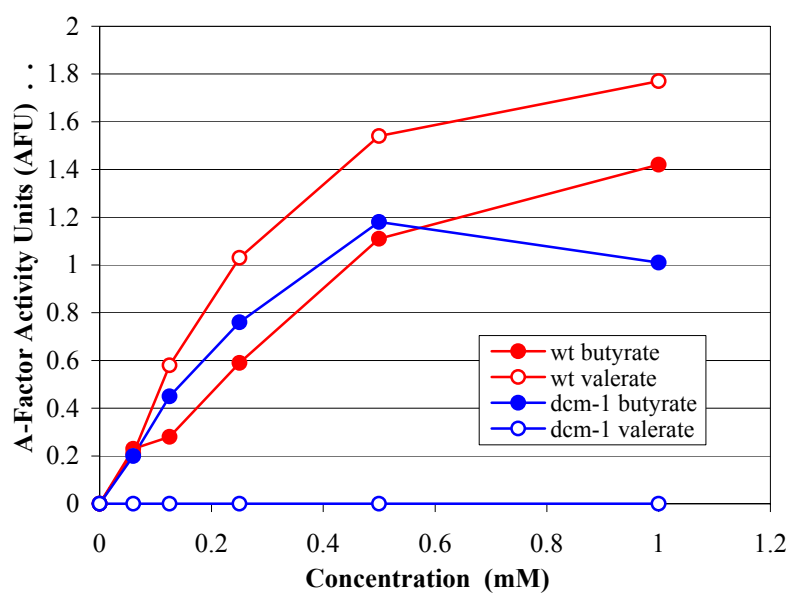


Figure 3.9 A-factor activities from the *dcm-1* strain compared with the wildtype in response to short chain fatty acid, butyrate and valerate. The *dcm-1* strain was not responsive to the stimulation by valerate, whereas very active in response to butyrate. The wild-type responded to both fatty acids.

DISCUSSION

A-signal is released into the development medium shortly after the wildtype *Myxococcus xanthus* is subjected to starvation. The mechanism by which *M. xanthus* cells respond to the chemically heterogeneous A-signal was previously unknown. The mechanism used by *M. xanthus* to respond to A-signal, a chemically heterogeneous mixture of several amino acids, has not been previously investigated. The hypothesis that A-factor response involves the same basic amino acid catabolism pathways used by *M. xanthus* for growth was tested using mutants with defects in specific amino acid catabolic enzymes. Mutants with defects in the BCAA, phenylalanine, or alanine catabolic pathways failed or severely impaired in the specific A-factor response to the corresponding amino acid. These mutants continued to respond to other A-factor amino acids that were tested. For example, a BCKAD mutant, that lacks an enzyme required for catabolism of all three BCAA, continued to respond to phenylalanine, proline and tyrosine. In addition, a number of new compounds were found to have A-factor activity (the ability to activate the A-factor dependent gene $\Omega 4521$) in this study. These compounds include pyruvate, the amino acids ornithine and citrulline, and fatty acids butyrate and valerate. All of these compounds (discussed below) are likely to serve as sources of carbon and energy for *M. xanthus* and, in this respect, to be similar to the common amino acids that have been shown to be important components of the A-factor released by cells. These results suggest that any compound that can serve as carbon and energy source, and be broken down into compounds feeding into the TCA cycle will have A-factor activity. These results also raise the possibility that compounds like fatty acids may be unrecognized components of the A-factor produced by *M. xanthus* cells. The fact that the compounds with A-factor activity are likely to be used to produce TCA cycle intermediates

also suggests these compounds are used to produce a common molecule that is used for the integration of the A-factor response to the different A-factor compounds and as the “signal” for the activation of the transcription of the A-factor dependent genes. Some of the protein components have been identified. These include SasA, SasN, SasP, SasR and SasS. The hypothetical signal molecule might bind to a protein component of the transcription regulatory system, such as ATP and H⁺.

A-signal processing requires A-factor amino acid catabolic pathways

Although A-signaling has been studied for many years by now, the research has been mostly focused on the A-signal production. It is essentially unknown at this time how A-factor is received and processed by the starving cells on a developmental medium. Based on the experimental results presented above, we offer a preliminary view into the A-signal reception side of the cellular processes.

First, the A-signal molecules are transported into cells actively via some transporters in the membrane. For example, *M. xanthus* genome database search shows that high affinity branched amino acid transporter genes exist. One of the apparent operon *livKHMGF* is at just over the 8.5 Mbp position on the chromosome. This coincides with the hypothesis that *M. xanthus* uses an amino acid transporter system with a $K_m \leq 2 \times 10^{-6}$ M (Manoil and Kaiser, 1982). The fact that *M. xanthus* grows on amino acids also suggests such amino acid transporters are present in *M. xanthus*.

Second, when signaling via dedicated molecules the signal molecules, do not need any special processing, mere their recognition by or binding to respective receptors is all that

needed to trigger an appropriate response from the receiving cell. For example, in *Pseudomonas* the autoinducer, an acyl-homoserine lactone, is recognized by its receptor LuxR and triggers activation of promoter for gene expression. The autoinducer is degraded when the cells enter into stationary phase, but that is for removal of the signal, not for any part of the signal transduction: signal generation, transmission, reception or response. The A-signal in *Myxococcus xanthus* is unique because 1) the signal mediator is a large set of molecules, not a single or a limited few dedicated molecules; 2) the signal molecules are the carbon and energy source for the cell. Since A-signal in *Myxococcus xanthus* is mediated by such a cohort of different molecules, it is hard to imagine any single or a limited few cellular apparatus to bind to or recognize all of them and sum up the respective activities to trigger the A-signal response. As we have demonstrated here with experimental results, once A-factor molecules get into developing cells, they have to be processed in a special way, in addition to the normal signal transduction processes.

As stated earlier, an extra high concentration of any A-factor constituents does not necessarily increase the strength of the A-signal because there is a maximum possible A-factor activity the cells can generate from a specific A-factor constituent. This probably reflects the rate at which the A-factor molecules are degraded inside the cell. The limit on a single A-signal mediator's contribution to the A-signal response seems to be present to all A-signal molecules, albeit to different levels obviously. This is represented by the tendency for the induced A-signal activity to level off or drop above certain concentration of the inducer. This is consistent with the proposed requirement of an A-factor degradation step in the A-factor response generation process.

A-signal response is activated by the A-factor catabolites (as carbon and energy sources)

The A-signal processing stage in *M. xanthus* involves a set of degradation pathways. These degradation pathways are of two types. One type is specific for each signal molecule. For example, each amino acid has its own specific degradation pathway, and different fatty acids have their special degradation pathways. The other type is the central / general metabolic pathway: the tricarboxylic acid (TCA) cycle and the glyoxylate cycle / bypass. The endproducts from the specific degradation pathways feed into the central metabolic pathways. The overall results from the signal processing stage are the production of NAD(P)H, FADH₂, H⁺, acetyl-CoA, oxaloacetate, and ATP. This probably allows a low level of biosynthesis to provide needed products for development, at the same time to exert control of the developmental program. Since this developmental process proceeds under the starvation condition, the cells maintain a high concentration of cAMP and ppGpp. These two molecules are known to be global regulators in bacteria (Johansson *et al.*, 2000). The A-signal response must be activated by one or more of these degradation endproducts in conjunction with the global regulators: ppGpp and cAMP. Among these catabolites, however, NADH and FADH₂ production seems to be not critical for the A-signal response. A careful examination of the BCAA degradation reactions (Figure 3.4) will discover that two molecules of NADH and one molecule of FADH₂ are generated for degrading each isoleucine before it reaches the propionyl-CoA carboxylase. In other words, the *dcm-1* mutant does not lack NADH and FADH₂ during its development. Therefore, the A-signal response defect observed in the *dcm-1* mutant suggests that ATP or H⁺ induced membrane voltage potential may be the activator for the A-signal dependent genes. For example, in *Bacillus subtilis*, the anti- σ SpoIIAB binds sporulation-specific σ^F in the presence of ATP, taking σ^F away from promoters, or binds the anti-anti- σ SpoIIAA in the presence of ADP,

activating the σ^F (Shu *et al.*, 2004). At the same time, still in *Bacillus subtilis*, Na⁺/H⁺ antiporter malfunction eliminates sporulation (Kosono *et al.*, 2004).

The glyoxylate cycle has to be involved because first, it is usually activated in bacteria during starvation to more efficiently utilize the carbon source. Second, it helps explain how the ketogenic amino acids such as leucine and fatty acids can be processed into the same endproducts, and used for the same function as the other A-factor amino acids. The glyoxylate cycle converts acetyl-CoA, and its precursors such as acetate, and acetoacetate from ketogenic processes to succinate, which furnishes the TCA cycle (Figure 3.10). This essentially makes all the known A-factor molecules equally well suited for gluconeogenesis during development. Third, homologs for glyoxylate specific enzymes isocitrate lyase and malate synthase apparently exist in the *M. xanthus* genome at 8.26 Mbp position, in an apparent operon of *aceBA* (Fig. 1.10). Fourth, previous reports show that the key enzymes isocitrate lyase and malate synthase are activated during the glycerol induced sporulation process (Bland *et al.*, 1971; Watson and Dworkin, 1968) and inactivated as the spores matured (Olowski and White, 1974).

The knowledge of the A-signal response mechanism provides the first opportunity to synthesize a view on the A-signal transduction system, including both A-signal production and response. In the following I will first summarize what is known in the literature about the A-signaling process in *M. xanthus* development, then present my view and produce an A-signaling model.

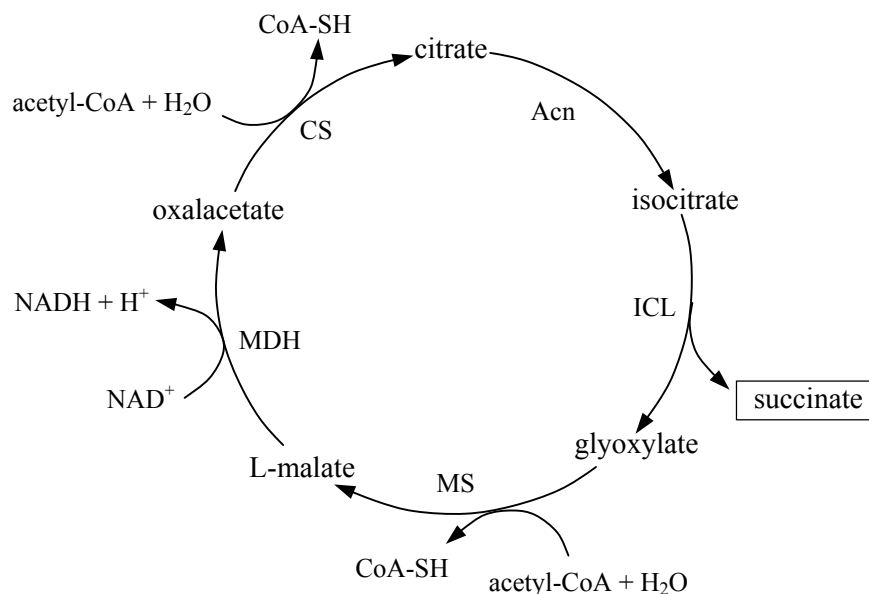


Fig. 3.10 The glyoxylate cycle CS, citrate synthase; Acn, aconitase; ICL, isocitrate lyase; MS, malate synthase; MDH, malate dehydrogenase.

A-factor production genes, *asgA*, *asgB*, and *asgC*, are not regulated by stringent response

Since *M. xanthus* development (fruiting body formation and sporulation therein) is triggered by starvation, the highest level of regulation in response to starvation is stringent response with increased level of alarmone, guanosine-5'-(tri)di-3'-diphosphate [(p)ppGpp, also known as a secondary messenger]. The alarmone mediated stringent response is quite universal in eubacteria (Mittenhuber, 2001; Artsimovitch *et al.*, 2004). Indeed, *M. xanthus* does use this mechanism for regulating the initiation of development (Manoil and Kaiser, 1980; Harris *et al.*, 1998). However, the relationship between the stringent response and development, and the A-signaling process in particular, seems to be a bit confusing and

worth a close examination.

There are five genes reported as involved in A-signal production, *asgA*, *asgB*, *asgC*, *asgD*, and *asgE*¹, respectively encoding a sensor for a two-component signal transduction system, a transcriptional factor, the major sigma factor *rpoD*, another sensor for a different two-component signal transduction system, and a cytosine deaminase. Because of the way a standard development assay is set up, *M. xanthus* cells are washed to remove all nutrient, resuspended in a starvation medium, and to initiate a development assay without any carbon and energy source. The cells obviously will go through a stringent response process. However, the stringent response, a dramatic cellular process, does not seem to have any effect on the expression of the first three, *asgA*, *asgB*, *asgC*. Their expression levels are almost constant during vegetative growth and development states (Plamann *et al.*, 1995; Plamann *et al.*, 1994; Davies *et al.*, 1995). Since cells are washed before development assay, the initial level of A-signal in the development medium must be zero. There must be a process for the A-signal level to reach a threshold level, and then a steady state level. However, the A-signal level change does not appear to affect the expression levels of the

¹ Garza and colleagues (Garza *et al.*, 2000) reported an *atzB* homolog, called *asgE*, as required for A-signaling. However, a close look at the sequence found that the sequence is most highly homologous (Score 234 and E 4e⁻⁶⁰) to an *ssnA* homolog in *Pseudomonas fluorescens* PfO-1, which encodes a cytosine deaminase. Cytosine deaminase converts cytosine to uracil in the cytosine catalysis pathway. The gene *ssnA* is activated during stationary phase in *E. coli* (Yamada *et al.*, 1999). This gene is located at 1.37 Mbp on the physical map of *M. xanthus* (Figure 1.10). According to our analysis, the gene should be predicted as a cytosine deaminase. Previous experiments show that it is required for development (Garza *et al.*, 2000).

Details on the analysis of *asgE* sequence: data is available in the NCBI GenBank showing that this gene encodes a conserved 421-amino-acid long domain that is most highly homologous to the domain COG0402 (Score 171 E 2e⁻⁴³). This domain is characteristic of cytosine deaminase and related metal-dependent hydrolases, such as the protein SsnA from *Pseudomonas fluorescens* PfO-1. This homology between *AsgE* and SsnA is much higher (Score 234 E 4e⁻⁶⁰) than either between *AsgE* and *AtzA* or *AsgE* and *AtzB*. The similarity and identity between *AsgE* and SsnA are respectively 51% and 36%. Therefore the predicted function for *AsgE* is more likely to be cytosine deaminase and related metal-dependent hydrolases. Cytosine deaminase converts cytosine to uracil in the cytosine degradation pathway. Consequently, the change in *AsgE* function prediction implies a possibility that nucleic acids could be A-factor components too.

first three A-signal genes, either. All these observations suggest that the functions of proteins coded by these genes do not change when the cellular conditions are converted from vegetative state to development state. That is, they are always involved in producing / releasing the complex mixture in the extracellular environment, contrary to previous claims (*cf.* Harris *et al.*, 1998). Most importantly, this complex mixture is called A-signal during development.

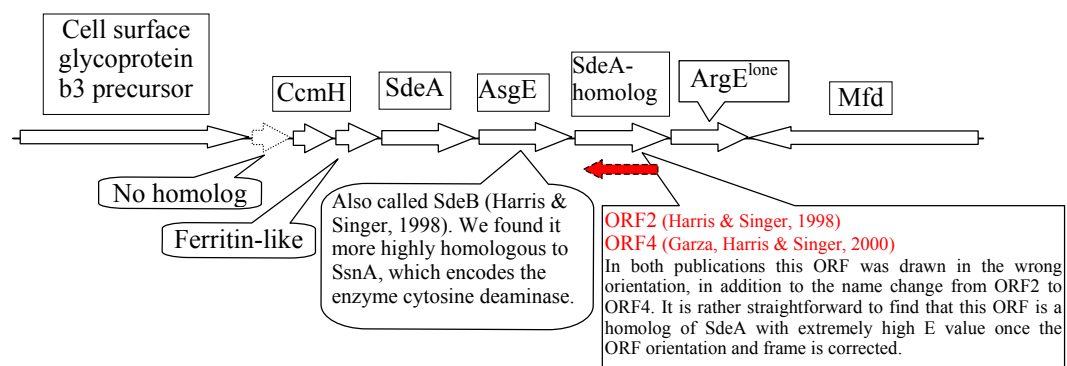


Fig. 3.11 The lone “*argE*” (*argE^{lone}*) site at 1.55 Mbp on the physical map. The *asgE* gene is in the same apparent operon as *argE^{lone}*.

In sharp contrast, the forth gene *asgD* is not detectable during the vegetative state, and dramatically increases expression after one hour of development. The protein AsgD is involved in both environmental sensing and intercellular signalling. However, this protein seems to be dispensible because the *asgD* mutant, while develops normally under total satrvation condition, develops poorly on a very low nutrient medium CF (Cho and Zusman, 1999). The AsgD protein slightly adjusts the developing cells’s sensitivity to the nutrient level.

Elevated (p)ppGpp concentration switches vegetative cells to the development state

The stringent response invariably results in accumulation of large amounts of (p)ppGpp in bacterial cells. In *M. xanthus*, Manoil and Kaiser (1980) suggested that raised (p)ppGpp concentrations serve as an intracellular signal, while Harris *et al.* (1998) called them signaling molecules, and a general starvation signal. Among the various nutritionally related regulatory systems in bacteria, stringent response apparently is a top level regulatory system, and controls other regulatory systems such as the catabolite repression / activation system (the cAMP-CRP modulon) (Johansson, 2000; Dorman, 2004), nitrogen regulated (Ntr) response (Reitzer, 2003), and Lrp (leucine-responsive regulatory protein) regulated systems (Reitzer, 2003). Since *Myxococcus xanthus* produces a large amount of (p)ppGpp in response to starvation, typical of normal stringent response, we could assume that the stringent response in *Myxococcus xanthus*, similar to other Eubacteria, results in a broad metabolic readjustment such as inhibition of ribosome, peptidoglycan, and phospholipid biosynthesis (Cashel, 1975; Cronan, 1978; Ishiguro and Ramey, 1978), and stimulation of intracellular proteolysis and the expression of biosynthesis operons such as *lac*, and *his* in *Escherichia coli* (Goldberg and St. John, 1976; Primakoff and Artz 1979; Stephens *et al.*, 1975). We infer from these findings that as starvation occurs *M. xanthus* cells immediately raises the (p)ppGpp concentration, and stops processes for growth and initiates processes to cope with the starvation situation.

Since (p)ppGpp is present in the growth phase (Manoil and Kaiser, 1980), the critical point for switching the *M. xanthus* to the development state is the elevated concentration of (p)ppGpp, that is, 10-12 times higher than that of exponentially growing phase level (Manoil and Kaiser, 1980; Harris *et al.*, 1998). Only under this elevated (p)ppGpp

condition, can cells be prepared for and receptive to developmental signals. This state of cell conditions, development state, is considered a unique state, different from that of growth phase. The elevated concentration of (p)ppGpp can be achieved by same mechanisms that increase (p)ppGpp in other organisms, such as general starvation, single (or more) amino acid starvation, phosphate starvation, glycerol-induction, and dimethyl sulfoxide-induction (Manoil and Kaiser, 1980).

Our experimental results showed that A-signal response was generated via the common metabolites in the TCA cycle; particularly, the central metabolite pyruvate is a potent A-factor activity inducer. Any one of a broad range of compounds that can be catabolized by *M. xanthus* could potentially be an A-signal mediator. Since expression of A-signal production genes, *asgA*, *asgB*, and *asgC*, are not regulated by the alarmone, ppGpp, it is unlikely that A-signal production would be regulated by ppGpp. In other words, potential A-factor is available in growth media, not regulated by stringent response, contrary to previous believe (Harris *et al.*, 1998). However, A-factor compounds (A-signal molecules) don't induce A-factor activity, or initiate development without an elevated intracellular concentration of (p)ppGpp as the preexisting condition. For example, *M. xanthus* grows on defined media such as A and A1, which contains enough known A-factor compounds (amino acids) that would trigger development if they were sensed as A-signal. Our results suggest that A-factor is not initiated by ppGpp as specialized compounds for A-signaling. The mode of action of A-signal suggests that any catabolizable carbon source could be potential A-signal molecule. It would be a mistake to believe that there is no A-factor in the growth media only because the cells don't sense the A-factor compounds as A-signal, or because the A-

factor does not activate developmental promoters such as Ω 4521. A-factor by definition (Kuspa *et al.*, 1992) is a set of chemicals, not how the cells react to them.

Consistent with our view is that the stringent response regulates many of the known developmental genes. First, *M. xanthus* has two genes homologous to *Escherichia coli* ATP-dependent protease *lon* gene: *lonV* and *lonD* (Tojo *et al.*, 1993). The LonD protease is required for development. Since *E. coli* Lon protease is activated by ppGpp-binding (Kuroda *et al.*, 2001), it is more likely that *M. xanthus* LonD is activated by ppGpp than by the initiation of A-factor production or A-signaling. A-factor could be provided if needed. But that would not rescue the developmental defects. Second, the *pho* operon in *E. coli* and *Bacillus subtilis* responds to the stringent response signal ppGpp by accumulating polyphosphate (polyP) (Rao *et al.*, 1998), which is critical for the DNA-binding activity and proteolytic activity of the ATP-dependent protease Lon in *E. coli* (Nomura *et al.*, 2004). Starvation for phosphate triggers stringent response in *M. xanthus* and accumulates huge amount of ppGpp (Manoil and Kaiser, 1980). Insertional mutation in the *phoRI* gene renders *M. xanthus* developmentally defective (Martinez-Canamero *et al.*, 2003). Comparing to the *E. coli* system, it is more likely that the inability to activate the protease LonD on development media due to lack of polyP causes the developmental defects in *phoRI*⁻ *M. xanthus* than lack of the initiation of A-factor production or A-signaling. A-factor could be provided if needed. But that would not rescue the developmental defects. This is because the cellular condition is not “readjusted” to the development state, and is not receptive to the developmental signals.

Since the *socE*¹ gene sequence cannot be verified by the *M. xanthus* genome database, further discussion will depend on where on the *M. xanthus* chromosome the *socE* sequence is to be found.

The window of A-factor concentration as A-signal

The minimal A-factor concentration is determined to be 10 μ M of any one of the six most effective A-factor amino acids, or an equimolar mixture thereof (10 μ M total) (Kuspa *et al.*, 1992b). The next question is whether there is an upper limit concentration for the A-factor constituents to be a useful A-signal. If there is, then what is the upper limit? In other words, we know that extra high concentration of a specific A-factor molecule does not necessarily increase the A-factor activity derived from that A-factor species. Some A-factors seem to even have a tendency to decrease their A-factor activity as their concentration increases over some limit, such as cysteine (limit undetectably low, always contributes negatively), tryptophan (limit 200 μ M) (Kuspa *et al.*, 1992a). But the question we address here is whether increasingly higher concentrations of an A-factor molecule would eventually cease its effect as an A-signal mediator.

To examine this limit, one has to provide the cells with high concentrations of amino acids, yet also to raise the (p)ppGpp concentration to ensure that if A-signal appears, the

¹ It is believed that the stringent response is regulated by the gene *socE* in *M. xanthus* (Crawford and Shimkets, 2000). But, the *socE* sequence has no homology to *relA* or *spoT*, which are known to regulate stringent response in bacteria. The gene *socE* has no homology to any sequence in the NCBI GenBank, except a very limited homology to a conserved domain COG3483 (TD01), tryptophan 2,3-dioxygenase (vermillion). According to the physical map of *M. xanthus* (He *et al.*, 1994), the *socE* (*soc537*, Rhie and Shimkets, 1989) gene lies at 1.01 Mbp position, which is covered by the contig506 (Fig. 1.10). However, the DNA sequence of the *socE* has no homology to any reasonable size of DNA fragment in the *Myxococcus xanthus* genome database (incomplete) at NCBI GenBank, except for a very limited homology (less than 100 bp) at about 80 kbp downstream the *socE* physical map location on the *M. xanthus* chromosome. Interestingly, it has very good protein homology to those translated from many cloning vectors.

development process is initiated. This is difficult to do because amino acids are the carbon and energy sources for *Myxococcus xanthus*. If there is a high concentration of amino acids in the medium, cells would not be sensing the situation as starvation, therefore they would not raise the (p)ppGpp concentration. Luckily, like in other prokaryotes, *M. xanthus* senses any single amino acid starvation as a general starvation (Dworkin 1963; Hemphill *et al.*, 1968; Manoil and Kaiser, 1980; Harris and Singer, 1998; Harris *et al.*, 1998; Ward *et al.*, 2000), and raises the concentration of (p)ppGpp as its response (Manoil and Kaiser, 1980; Harris *et al.*, 1998; Garza *et al.*, 2000). These experiments show that amino acid auxotrophs develop on chemically defined media, and artificially inducible *relA* gene also induce cells to initiate development on rich media. The upper limit for A-factor concentration seems to be beyond their concentration in the growth media. Although concentrations higher than they are in the media may still impose some negative affect, that is beyond our concern because we want to determine whether the A-factor concentration as high as in the growth media still functions as A-factor.

A model for the A-signaling process

The findings presented in this work substantiated the mechanism of A-signal transduction. However, there have been speculations about how *M. xanthus* development is regulated. For example, working on an alanine dehydrogenase mutant, *aldA*, which carries developmental defects, Ward and colleagues (Ward *et al.* 2000) suggested that a specialized regulatory mechanism, similar to the leucine-responsive regulatory protein, Lrp¹, might be the

¹ A search in the incomplete *M. xanthus* genome sequence database at NCBI found that there is an Lrp protein at around 0.5 Mbp position on the physical map (He *et al.*, 1994) (Fig. 1.10). This protein is 158-residues long, and referred to as a “putative transcriptional regulatory protein” (Galbis, 2001 [thesis, Universidad de Murcia, Murcia, Spain]). It has a 152-amino-acid Lrp domain homology with a score 126, E value $2e^{-30}$. This putative transcriptional regulatory protein has a 60% similarity and 39% identity with the leucine-responsive regulatory protein from *Sinorhizobium meliloti*.

regulatory component. However, even if the *aldA* mutant results can be explained by an Lrp mechanism, it is still hard to explain the other A-factor behavior in the mutants discussed in this chapter. For example, loss of branched chain keto acid dehydrogenase in *esg* may lead to the accumulation of branched chain keto acids and branched chain amino acids; loss of propionyl-CoA may lead to the accumulation of propionic acid and other odd chain fatty acids. It is difficult to imagine that there are responsive regulatory proteins for each and every one of these A-factor molecules. Therefore, the actual A-signal response must use a mechanism other than an Lrp-like system.

Our evidence shows that several mutations in amino acid degradation pathways abolish the A-signal response to the cognate amino acids. We also discovered that some short chain fatty acids and pyruvate have A-factor activity, and they too activate A-signal dependent regulation through their degradation pathways. In summary, this work signifies that the A-signaling process goes through the A-factor mediator degradation pathways and the central metabolic cycles: TCA and glyoxylate cycles.

Another set of A-signal regulation information comes from the *asg* suppressors studies. Some *asg* suppressors cause a moderate level of expression of the Ω 4521 promoter in both the vegetative state and in the development state. These mutants carry a mutation in the *sasS* gene (Yang and Kaplan, 1997). The SasS protein is a two-component (histidine kinase) sensor. SasS is a negative regulator for the Ω 4521 expression. The point mutation in the *sasS* gene is such that it causes unrestrained phosphate transfer to its downstream cognate, while the developmental stimulation is unaltered (Kaplan *et al.*, 1991). Some other *asg*

suppressors cause a high level expression of $\Omega 4521$ during the vegetative state. and express even more during the development state (Kaplan *et al.*, 1991). These mutants carry a mutation in the *sasN* gene (Xu *et al.*, 1998). The SasN protein is a negative regulator of the promoter $\Omega 4521$, and functions upstream of the SasS (Xu *et al.*, 1998). It has an excellent leading signal peptide. Still other *asg* suppressors show that the regulation of $\Omega 4521$ exists either downstream *sasS*, i.e. *sasR*¹, or upstream *sasS*, i.e. *sasP*. The *sasR* is a two-component receiver/regulator, presumably a cognate of SasS, and forming a complete two-component system with SasS.

This interaction profile of the three *sas* genes (*sasN*, *sasS*, and *sasR*) is reminiscent of an extracytoplasmic function (ECF) sigma factor system (Raivio *et al.*, 2001 for an excellent review on ECF; Browning *et al.*, 2003) (Fig. 3.12). In ECF terms, as shown in Figure 3.12, SasR is a σ factor, SasS an anti- σ factor, and SasN an antianti- σ factor. The mechanism might like this: the development signal is sensed by SasS, but SasN binds to the SasS preventing SasS from phosphorylating SasR, the ECF σ factor. If SasN is removed, e.g. in *sasN* mutants, SasS will activate (phosphorylate) the SasR, which in turn activates the expression of $\Omega 4521$. Since this sigma system is speicalized to respond to the extracytoplasmic conditions, the sigma factor (SasR) is called extracytoplasmic function sigma factor (ECF σ). The other component in this chain of interaction is the protein SasP, a positive regulator of $\Omega 4521$ expression, and operating upstream of SasS. Follwing the ECF

¹ The *sasR* and *sasP* were identified based on *sasA* that is an *rfbA* homolog, which encodes a component for an ABC-transporter (Gou *et al.*, 1996). This ABC-transporter is critical for the synthesis of an O-antigen on the *M. xanthus* cell surface. Loss of *sasA* leads to $\Omega 4521$ expression in the starving cells without A-signal. Based on the *sasA* phenotype, Guo and colleagues discovered transposon insertion suppressors of *sasA*, the *ssp* class, which essentially restores the wildtype patterns of $\Omega 4521$ expression to the *sasA* mutant (Guo *et al.*, 2000).

sigma factor system analogy, the SasP would be an antianti- σ factor, which inactivates the antisigma factor SasN. This analysis reveals the analogy between the *M. xanthus* ECF σ and the sporulation sigma factor σ^F of *Bacillus subtilis* (Shu *et al.*, 2004). Interestingly, we found additional supporting evidence, too: 1) SasN has an excellent predicted leading signal peptide with a potential cleavage site between the two alanine residues at position 35 and 36 (Bendtsen *et al.*, 2004). This shows the SasN is a member of the periplasmic space. 2) SasP¹ is a secreted protein too. 3) Since the suppressors of *sasA* gene, the *ssp* class, has saturated the genome (Guo *et al.*, 2000), we interpret it as there is no unidentified gene product to inactivate SasN other than SasP. Therefore the representation in the Figure 3.12 reflects the actual relationship.

The most intriguing part of the Sas-related data is how the O-antigen polysaccharide could be explained in relation to the signaling results. We suggest that the polysaccharide part from the O-antigen bends back on the signal sensors such as SasS or its related elements and inhibits the sensor. When on stable medium with a solid surface and in high cell density, the polysaccharides function as a tactile sensor (Lee *et al.*, 1995), and are engaged with the polysaccharides from the neighboring cells. Therefore there is much less chance for polysaccharides to fall back, touch and inhibit the A-signal sensor on the cell surface. This would require a considerable cell density to keep polysaccharides engaged in tactile “handshaking” and keep the A-signaling channel open. The fact that some of the suppressors of the *sas* genes (the *ssp* series) are involved in extracytoplasmic

¹ But its start codon has to extend 13 codons upstream from the previously predicted first ATG (Guo *et al.*, 2000) to the second GTG in the open reading frame. After this extension, SasP carries a signal peptide with a potential cleavage site between the alanine and glutamate residues at position 23 and 24 (Bendtsen *et al.*, 2004).

polysaccharide synthesis (Guo *et al.*, 2000) further supports this interpretation. If this interpretation should be proven, we would have a much better integrated understanding of the way the cells sense and interact with each other and the environment.

Why the *sas* genes are discussed in here? First, they are clearly involved in the A-signal transduction. Second, they were thought as involved in A-signal response. Our experimental results show that A-signal response requires the catabolic system to generate carbon and energy. The *sas* genes are clearly not part of that. The *sas* genes are clearly directly involved in the gene regulation system. It is hard to imagine an A-signaling system without them. However, there could be two separate but parallel processes in so-called A-signaling process: One mediated via A-signal molecules (amino acids) by the catabolic system primarily for carbon and energy, the other via extracytoplasmic environmental condition sensed by *sas* genes. These two processes could be related by assuming A-signal response is energy dependent and environmental condition dependent. The catabolic processes provide carbon and energy for proper response to the environmental (starvation) condition.

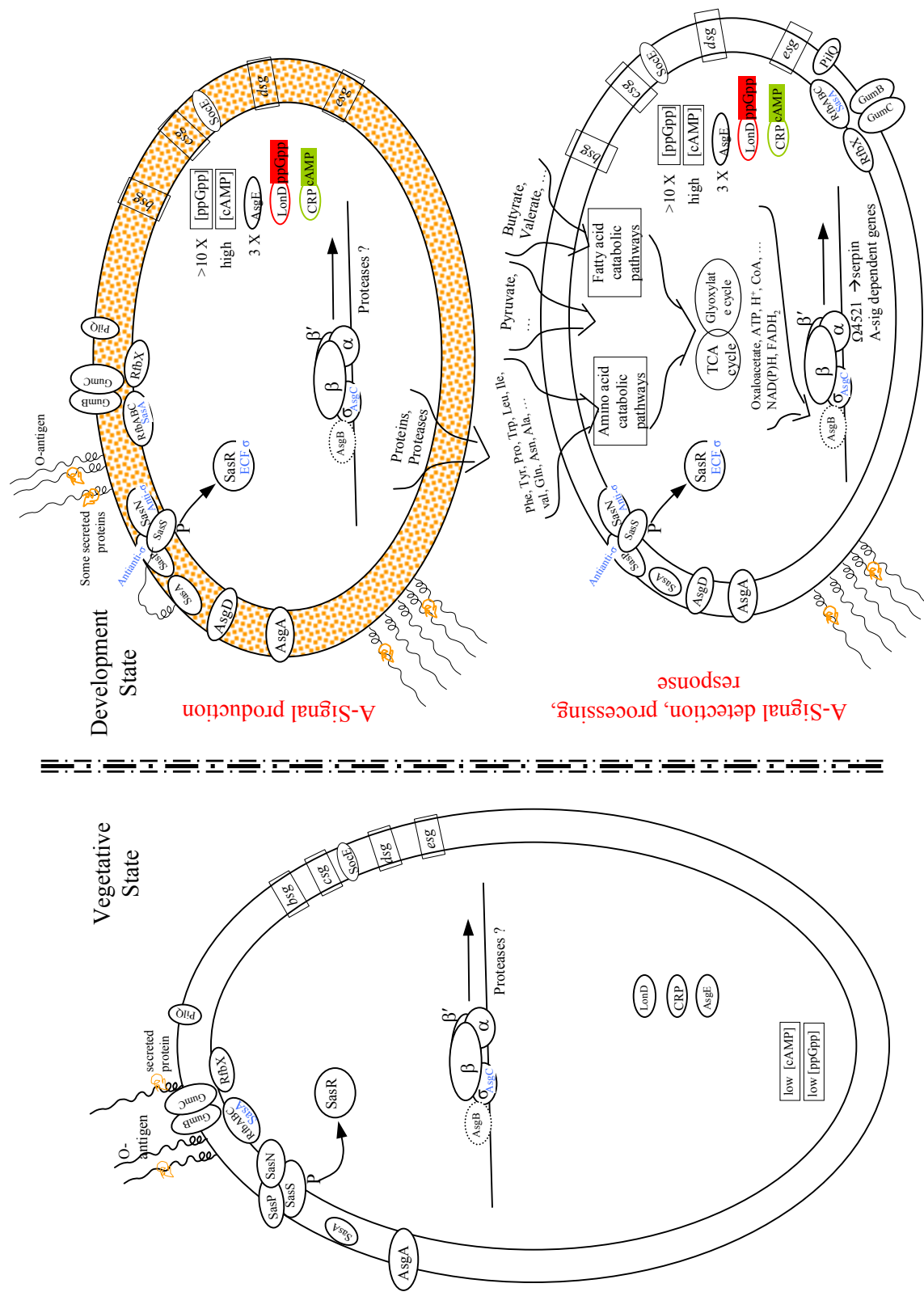


Figure 3.12 The A-signal model. See text for explanation.

REFERENCES

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. *J Mol Biol.* 215(3):403-10.
- Artsimovitch I, Patlan V, Sekine S, Vassilyeva MN, Hosaka T, Ochi K, Yokoyama S, Vassilyev DG.** 2004. Structural basis for transcription regulation by alarmone ppGpp. *Cell.* 117(3):299-310.
- Bartholomeusz GA.** 1998. Study of the role of the *M. xanthus* *esg* locus in development and lipid biosynthesis. Dissertation. University of Oklahoma, Norman, Oklahoma, USA.
- Barton GJ, Sternberg MJ.** 1987. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J Mol Biol.* 198(2):327-37.
- Bellenger K, Ma X, Shi W, Yang Z.** 2002. A CheW homologue is required for *Myxococcus xanthus* fruiting body development, social gliding motility, and fibril biogenesis. *J Bacteriol.* 184(20):5654-60.
- Behmlander RM, Dworkin M.** 1991. Extracellular fibrils and contact-mediated cell interactions in *Myxococcus xanthus*. *J Bacteriol.* 173(24):7810-20.
- Behmlander RM, Dworkin M.** 1994. Biochemical and structural analyses of the extracellular matrix fibrils of *Myxococcus xanthus*. *J Bacteriol.* 176 (20): 6295–6303.
- Bensmail L, Quillet L, Petit F, Barray S, Guespin-Michel JF.** 1998. Regulation of the expression of a gene encoding beta-endoglucanase secreted by *Myxococcus xanthus* during growth: role of genes involved in developmental regulation. *Res Microbiol.* 149(5):319-26.
- Bendtsen JD, Nielsen H, Heijne G, Brunak S.** 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, 340:783-795.
- Bonner PJ, Shimkets LJ.** 2001. Piecing together a puzzling pathway: new insights into C-signaling. *Trends Microbiol.* 9(10):462-4.

Boundy-Mills, K.L., De Souza, M.L., Mandelbaum, R.T., Wackett, L.P., Sadowsky, M. 1997. The *atzB* gene of *Pseudomonas* sp. strain ADP encodes the second enzyme of a novel atrazine degradation pathway. *Appl Environ Microbiol* 63: 916-923.

Bowden MG, Kaplan HB. 1998. The *Myxococcus xanthus* lipopolysaccharide O-antigen is required for social motility and multicellular development. *Mol Microbiol.* 30(2):275-84.

Browning DF, Whitworth DE, Hodgson DA. 2003. Light-induced carotenogenesis in *Myxococcus xanthus*: functional characterization of the ECF sigma factor CarQ and antisigma factor CarR. *Mol Microbiol.* 48(1):237-51.

Brenner M., Garza AG., Singer M. 2004. *nsd*, a Locus That Affects the *Myxococcus xanthus* Cellular Response to Nutrient Concentration. *J Bacteriol.* 186(11):3461-71.

Bretscher AP, Kaiser D. 1978. Nutrition of *Myxococcus xanthus*, a fruiting myxobacterium. *J Bacteriol.* 133(2):763-8.

Caberoy NB, Welch RD, Jakobsen JS, Slater SC, Garza AG. 2003. Global mutational analysis of NtrC-like activators in *Myxococcus xanthus*: identifying activator mutants defective for motility and fruiting body development. *J Bacteriol.* 185(20):6083-94.

Campagne F. 2000. Clustalnet: the joining of Clustal and CORBA. *Bioinformatics.* 16(7):606-12.

Campos JM, Zusman DR. 1975. Regulation of development in *Myxococcus xanthus*: effect of 3':5'-cyclic AMP, ADP, and nutrition. *Proc Natl Acad Sci U S A.* 72(2):518-22.

Cashel, M. 1975. Regulation of bacterial ppGpp and pppGpp. *Annu. Rev. Microbiol.* 29:301-318.

Cashel, M., Gentry, D. R., Hernandez, V. J. & Vinella, D. 1996. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*, ed. Neidhardt, F. C. (Am. Soc. Microbiol., Washington, DC), 2nd Ed., pp. 1458-1496.

- Chain P, Kurtz S, Ohlebusch E, Slezak T.** 2003. An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. *Brief Bioinform.* 4(2):105-23.
- Cheng YL, Kalman LV, Kaiser D.** 1994. The *dsg* gene of *Myxococcus xanthus* encodes a protein similar to translation initiation factor IF3. *J Bacteriol.* 176(5):1427-33.
- Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD.** 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31(13):3497-500.
- Cho, K. & Zusman, D.R.** 1999. *AsgD*, a new two-component regulator required for A-signaling and nutrient sensing during early development of *Myxococcus xanthus*. *Mol Microbiol* 34: 268-281.
- Crawford EW Jr, Shimkets LJ.** 2000. The stringent response in *Myxococcus xanthus* is regulated by *SocE* and the *CsgA* C-signaling protein. *Genes Dev.* 14(4):483-92.
- Cronan, J.** 1978. Molecular biology of bacterial membrane lipids. *Annu. Rev. Biochem.* 47:163-189.
- Cummings L, Riley L, Black L, Souvorov A, Resenchuk S, Dondoshansky I, Tatusova T.** 2002. Genomic BLAST: custom-defined virtual databases for complete and unfinished genomes. *FEMS Microbiol Lett.* 216(2):133-8.
- Davies DG, Parsek MR, Pearson JP, Iglewski BH, Costerton JW, Greenberg EP.** 1998. The involvement of cell-to-cell signals in the development of a bacterial biofilm. *Science.* 280(5361):295-8.
- Davis, J.M., Mayor, J., Plamann, L.** 1995. A missense mutation in *rpoD* results in an A-signaling defect in *Myxococcus xanthus*. *Mol Microbiol* 18: 943-952.
- Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C.** 1978. A model for evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3 (Dayhoff, M.O.,

ed.), pp 345-352, National Biochemical Research Foundation, Washington DC.

De Souza, M.L., Sadowsky, M.J., Wackett, L.P. 1996. Atrazine chlorohydrolase from *Pseudomonas* sp. strain ADP: gene sequence, enzymatic purification and protein characterization. *J. Bacteriol.* 178:4894-4900.

Downard JS. 1988. Tn5-mediated transposition of plasmid DNA after transduction to *Myxococcus xanthus*. *J. Bacteriol.* 170(10):4939-41.

Downard, J., Ramaswamy, S.V., Kil, K.-S. 1993. Identification of *esg*, a genetic locus involved in cell-cell signaling during *Myxococcus xanthus* development. *J. Bacteriol.* 175:7762-7770.

Downard J, Toal D. 1995. Branched-chain fatty acids: the case for a novel form of cell-cell signaling during *Myxococcus xanthus* development. *Mol Microbiol.* 16(2):171-5.

Downard J, Ramaswamy SV, Kil KS. 1993. Identification of *esg*, a genetic locus involved in cell-cell signaling during *Myxococcus xanthus* development. *J. Bacteriol.* 175(24):7762-70.

Dworkin, M. 1963. Nutritional regulation of morphogenesis in *Myxococcus xanthus*. *J. Bacteriol.* 86:67-72.

Dworkin, M. 1996. Recent advances in the social and developmental biology of Myxobacteria. *Microbiol. Rev.* 60: 70-102.

Eberl L. 1999. N-acyl homoserinelactone-mediated gene regulation in gram-negative bacteria. *Syst Appl Microbiol.* 22(4):493-506.

Elmi A, Idahl LA, Sehlin J. 2000. Relationships between the Na(+)/K(+) pump and ATP and ADP content in mouse pancreatic islets: effects of meglitinide and glibenclamide. *Br. J. Pharmacol.* 131(8):1700-6.

Falquet L, Bordoli L, Ioannidis V, Pagni M, Jongeneel CV. 2003. Swiss EMBnet node web server. *Nucleic Acids Res.* 31(13):3782-3.

Federle MJ, Bassler BL. 2003. Interspecies communication in bacteria. *J Clin Invest.* 112(9):1291-9.

Feng, DF. and Doolittle, RF. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol.* 25(4):351-60.

Ferreira SH and Vane JR. 1967. Prostaglandins: their disappearance from and release into the circulation. *Nature (Lond)* 216: 868-873.

Galbis,M.L. 2001. Global action of CarD on transcription regulation in *Myxococcus xanthus*. In: *Thesis, Department of Genetica y Microbiologia*, Universidad de Murcia, Murcia, Spain.

Garza AG, Harris BZ, Pollack JS, Singer M. 2000. The *asgE* locus is required for cell-cell signaling during *Myxococcus xanthus* development. *Mol. Microbiol.* 35(4):812-24.

Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. 2003. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31(13):3784-8.

Geng Y, Yang Z, Downard J, Zusman D, Shi W. 1998. Methylation of FrzCD defines a discrete step in the developmental program of *Myxococcus xanthus*. *J. Bacteriol.* 180(21):5765-8.

George RA, Heringa J. 2002. Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins.* 48(4):672-81.

Gill RE, Karlok M, Benton D. 1993. *Myxococcus xanthus* encodes an ATP-dependent protease which is required for developmental gene transcription and intercellular signaling. *J. Bacteriol.* 175(14):4538-44.

Gilbert, D. 2000. Free software in molecular biology for Macintosh and MS Windows computers. In: *Bioinformatics: methods and protocols.* (ed Misener, Stephen and Krawetz, Stephen A.) Series: Methods in Molecular biology. Vol 132,pp149-184

Goldberg, A., and A. St. John. 1976. Intracellular protein degradation in mammal and bacterial cells: part 2. Annu. Rev. Biochem. 45:747-803.

Guo D, Bowden MG, Pershad R, Kaplan HB. 1996. The *Myxococcus xanthus* rfbABC operon encodes an ATP-binding cassette transporter homolog required for O-antigen biosynthesis and multicellular development. J. Bacteriol. 178(6):1631-9.

Hagen, D.C., Bretscher, A.P., Kaiser, D. 1978. Synergism between morphogenetic mutants of *Myxococcus xanthus*. Dev. Biol. 64: 284-296.

Hager E, Tse H, Gill RE. 2001. Identification and characterization of *spdR* mutations that bypass the BsgA protease-dependent regulation of developmental gene expression in *Myxococcus xanthus*. Mol. Microbiol. 39(3):765-80.

Harris BZ, Kaiser D, Singer M. 1998. The guanosine nucleotide (p)ppGpp initiates development and A-factor production in *Myxococcus xanthus*. Genes Dev. 12(7):1022-35.

He Q, Chen H, Kuspa A, Cheng Y, Kaiser D, Shimkets LJ. 1994. A physical map of the *Myxococcus xanthus* chromosome. Proc Natl Acad Sci U S A. 91(20):9584-7.

Hemphill H. E., Zahler S. A. 1968. Nutritional induction and suppression of fruiting in *Myxococcus xanthus* FBa. J. Bacteriol. 95:1018-1023.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 89(22):10915-9.

Higgins DG, Thompson JD, Gibson TJ. 1996. Using CLUSTAL for multiple sequence alignments. Methods Enzymol. 266:383-402.

Hokamp K, Shields DC, Wolfe KH, Caffrey DR. 2003. Wrapping up BLAST and other applications for use on Unix clusters. Bioinformatics. 19(3):441-2.

Hvorup RN, Winnen B, Chang AB, Jiang Y, Zhou XF, Saier MH Jr. 2003. The multidrug/oligosaccharidyl-lipid/polysaccharide (MOP) exporter superfamily. Eur. J.

Biochem. 270(5):799-813.

Ibrahim HA, El-Meligi AA, Abdel-Hamid M, Elhendy A. 2004. Relations between von Willebrand factor, markers of oxidative stress and microalbuminuria in patients with type 2 diabetes mellitus. *Med. Sci. Monit.* 10(3):CR85-9.

Ishiguro, E. E., and W. D. Ramey. 1978. Involvement of the *relA* gene product and feedback inhibition in the regulation of UDP-N-acetyl-muramyl-peptide synthesis in *Escherichia coli*. *J. Bacteriol.* 135:766-774.

Johansson J, Balsalobre C, Wang SY, Urbonaviciene J, Jin DJ, Sonden B, Uhlin BE. 2000. Nucleoid proteins stimulate stringently controlled bacterial promoters: a link between the cAMP-CRP and the (p)ppGpp regulons in *Escherichia coli*. *Cell.* 102(4):475-85.

Shu JC, Clarkson JR, Yudkin MD. 2004. Studies of SpoIIAB mutant proteins elucidate the mechanisms that regulate the developmental transcription factor sigma F in *Bacillus subtilis*. *Biochem. J.* (2004 Aug 5;Pt. [Epub ahead of print])

Kaiser D. 2003. Coupling cell movement to multicellular development in myxobacteria. *Nat. Rev. Microbiol.* 1(1):45-54.

Kaiser, D. & Losick, R. 1993. How and why bacteria talk to each other. *Cell.* 73: 873-885.

Kalman LV, Cheng YL, Kaiser D. 1994. The *Myxococcus xanthus* *dsg* gene product performs functions of translation initiation factor IF3 in vivo. *J. Bacteriol.* 176(5):1434-42.

Kim SH, Ramaswamy S, Downard J. 1999. Regulated exopolysaccharide production in *Myxococcus xanthus*. *J. Bacteriol.* 181(5):1496-507.

Kim SK, Kaiser D. 1991. C-factor has distinct aggregation and sporulation thresholds during *Myxococcus* development. *J. Bacteriol.* 173(5):1722-8.

Kim SK. & Kaiser D. 1992. Control of cell density and Pattern by intercellular signaling in *Myxococcus* development. *Annu. Rev. Microbiol.* 46:117-39.

- Kimball SR, Jefferson LS.** 2004. Molecular mechanisms through which amino acids mediate signaling through the mammalian target of rapamycin. *Curr. Opin. Clin. Nutr. Metab. Care.* 7(1):39-44.
- Kimura Y, Sato R, Mimura K, Sato M.** 1997. Propionyl coenzyme A carboxylase is required for development of *Myxococcus xanthus*. *J. Bacteriol.* 179(22):7098-102.
- Kirby JR, Zusman DR.** 2003. Chemosensory regulation of developmental gene expression in *Myxococcus xanthus*. *Proc. Natl. Acad. Sci. U S A.* 100(4):2008-13.
- Kohli DK, Bachhawat AK.** 2003. CLOURE: Clustal Output Reformatter, a program for reformatting ClustalX/ClustalW outputs for SNP analysis and molecular systematics. *Nucleic Acids Res.* 31(13):3501-2.
- Korf I.** 2003. Serial BLAST searching. *Bioinformatics.* 19(12):1492-6.
- Kosono S, Asai K, Sadaie Y, Kudo T.** 2004. Altered gene expression in the transition phase by disruption of a Na⁺/H⁺ antiporter gene (*shaA*) in *Bacillus subtilis*. *FEMS Microbiol. Lett.* 232(1):93-9.
- Kroos, L. & Kaiser, D.** 1984. Construction of Tn5*lac*, a transposon that fuses *lacZ* expression to exogenous promoters, and its introduction into *Myxococcus xanthus*. *Proc. Natl. Acad. Sci. USA* 81: 5816-20.
- Kroos, L. & Kaiser, D.** 1987. Expression of many developmentally regulated genes in *Myxococcus* depends on a sequence of cell interactions. *Genes Dev.* 1: 840-854.
- Kroos, L., Kuspa, A., Kaiser, D.** 1986. A global analysis of developmentally regulated genes in *Myxococcus xanthus*. *Dev. Biol.* 117: 252-266.
- Kuroda A, Nomura K, Ohtomo R, Kato J, Ikeda T, Takiguchi N, Ohtake H, Kornberg A.** 2001. Role of inorganic polyphosphate in promoting ribosomal protein degradation by the Lon protease in *E. coli*. *Science.* 293(5530):705-8.

- Kuspa, A. & Kaiser, D.** 1989. Genes required for developmental signaling in *Myxococcus xanthus*: three *asg* loci. J. Bacteriol. 171: 2762-72.
- Kuspa, A., Kroos, L., Kaiser, D.** 1986. Intercellular signaling is required for developmental gene expression in *Myxococcus xanthus*. Dev. Biol. 117: 267-76.
- Kuspa, A., Plamann, L., Kaiser, D.** 1992a. Identification of heat-stable A-factor from *Myxococcus xanthus*. J. Bacteriol. 174: 3319-26.
- Kuspa, A., Plamann, L., Kaiser, D.** 1992b. A-signaling and the cell density requirement for *Myxococcus xanthus* development. J. Bacteriol. 174: 7360-69.
- Kuspa A, Vollrath D, Cheng Y, Kaiser D.** 1989. Physical mapping of the *Myxococcus xanthus* genome by random cloning in yeast artificial chromosomes. Proc. Natl. Acad. Sci. U S A. 86(22):8917-21.
- Lancero H, Brofft JE, Downard J, Birren BW, Nusbaum C, Naylor J, Shi W, Shimkets LJ.** 2002. Mapping of *Myxococcus xanthus* social motility *dsp* mutations to the *dif* genes. J. Bacteriol. 184(5):1462-5.
- LaRossa, R., Kuner, J., Hagen, D., Manoil, C., Kaiser, D.** 1983. Developmental cell interactions of *Myxococcus xanthus*: analysis of mutants. J. Bacteriol. 153: 1394-404.
- Lee BU, Lee K, Mendez J, Shimkets LJ.** 1995. A tactile sensory system of *Myxococcus xanthus* involves an extracellular NAD(P)(+)-containing protein. Genes Dev. 9(23):2964-73.
- Lee K, Shimkets LJ.** 1996. Suppression of a signaling defect during *Myxococcus xanthus* development. J. Bacteriol. 178(4):977-84.
- Li S, Lee BU, Shimkets LJ.** 1992. *csgA* expression entrains *Myxococcus xanthus* development. Genes Dev. 6(3):401-10.

Li Y, Sun H, Ma X, Lu A, Lux R, Zusman D, Shi W. 2003. Extracellular polysaccharides mediate pilus retraction during social motility of *Myxococcus xanthus*. Proc. Natl. Acad. Sci. USA. 100(9):5443–8.

Lipman DJ, Altschul SF, Kececioglu JD. 1989. A tool for multiple sequence alignment. Proc. Natl. Acad. Sci. USA. 86(12):4412-5.

Lobedanz S. and Sogaard-Andersen L. 2003. Identification of the C-signal, a contact-dependent morphogen coordinating multiple developmental responses in *Myxococcus xanthus*. Genes Dev. 17:2151-2161. (Published online before print August 15, 2003.)

Lussier M, White AM, Sheraton J, di Paolo T, Treadwell J, Southard SB, Horenstein CI, Chen-Weiner J, Ram AF, Kapteyn JC, Roemer TW, Vo DH, Bondoc DC, Hall J, Zhong WW, Sdicu AM, Davies J, Klis FM, Robbins PW, Bussey H. 1997. Large scale identification of genes involved in cell surface biosynthesis and architecture in *Saccharomyces cerevisiae*. Genetics. 147(2):435-50.

Manoil, C., and Kaiser, D. 1980a. Accumulation of guanosine tetraphosphate and guanosine pentaphosphate in *Myxococcus xanthus* during starvation and myxospore formation. J. Bacteriol. 141(1):297-304.

Manoil C. and Kaiser D. 1980b. Guanosine pentaphosphate and guanosine tetraphosphate accumulation and induction of *Myxococcus xanthus* fruiting body development. J. Bacteriol. 141(1):305-315.

Martinez-Canamero M, Munoz-Dorado J, Farez-Vidal E, Inouye M, Inouye S. 1993. Oar, a 115-kilodalton membrane protein required for development of *Myxococcus xanthus*. J. Bacteriol. 175(15):4756-63.

Martinez-Canamero M, Ortiz-Codorniu C, Extremera AL, Munoz-Dorado J, Arias JM. 2003. *phoR1*, a gene encoding a new histidine protein kinase *Myxococcus xanthus*. Antonie Van Leeuwenhoek. 83(4):361-8.

Mathog DR. 2003. Parallel BLAST on split databases. Bioinformatics. 19(14):1865-6.

- Mayer ML, Armstrong N.** 2004. Structure and function of glutamate receptor ion channels. *Annu. Rev. Physiol.* 66:161-81.
- McBride MJ, Zusman DR.** 1989. Trehalose accumulation in vegetative cells and spores of *Myxococcus xanthus*. *J. Bacteriol.* 171(11):6383-6.
- Mittenhuber G.** 2001. Comparative genomics and evolution of genes encoding bacterial (p)ppGpp synthetases/hydrolases (the Rel, RelA and SpoT proteins). *J. Mol. Microbiol. Biotechnol.* 3(4):585-600.
- Morris SM Jr.** 2004. Recent advances in arginine metabolism. *Curr. Opin. Clin. Nutr. Metab. Care.* 7(1):45-51.
- Mukamolova GV, Kaprelyants AS, Young DI, Young M, Kell DB.** 1998. A bacterial cytokine. *Proc. Natl. Acad. Sci. USA.* 95(15):8916-21.
- Needleman, S.B. and Wunsch, C.D.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48(3):443-53.
- Nomura K, Kato J, Takiguchi N, Ohtake H, Kuroda A.** 2004. Effects of inorganic polyphosphate on the proteolytic and DNA-binding activities of Lon in *Escherichia coli*. *J. Biol. Chem.* 279(33):34406-10.
- Nomura T, Lu R, Pucci ML, Schuster VL.** 2004. The two-step model of prostaglandin signal termination: in vitro reconstitution with the prostaglandin transporter and prostaglandin 15 dehydrogenase. *Mol. Pharmacol.* 65(4):973-8.
- Orlowski M, White D.** 1974. Inactivation of isocitrate lyase during myxospore development in *Myxococcus xanthus*. *J. Bacteriol.* 118(1):96-102.
- Osorio EC, de Souza JE, Zaiats AC, de Oliveira PS, de Souza SJ.** 2003. pp-BLAST: a "pseudo-parallel" BLAST. *Braz. J. Med. Biol. Res.* 36(4):463-4.
- Paquola AC, Machado AA, Reis EM, Da Silva AM, Verjovski-Almeida S.** 2003. Zerg: a very fast BLAST parser library. *Bioinformatics.* 19(8):1035-6.

Paulsen IT, Beness AM, Saier MH Jr. 1997. Computer-based analyses of the protein constituents of transport systems catalysing export of complex carbohydrates in bacteria. *Microbiology*. 143 (Pt 8):2685-99.

Peabody CR, Chung YJ, Yen MR, Vidal-Ingigliardi D, Pugsley AP, Saier MH Jr. 2003. Type II protein secretion and its relationship to bacterial type IV pili and archaeal flagella. *Microbiology*. 149(Pt 11):3051-72.

Plamann, L., Kuspa, A., Kaiser, D. 1992. Proteins that rescue A-signal defective mutants of *Myxococcus xanthus*. *J. Bacteriol.* 174: 3311-8.

Plamann, L., Davis, J.M., Cantwell, B., Mayor, J. 1994. Evidence that *asgB* encodes a DNA- binding protein essential for growth and development of *Myxococcus xanthus*. *J. Bacteriol.* 176: 2013-20.

Plamann, L., Li, Y., Cantwell, B., Mayor, J. 1995. The *Myxococcus xanthus asgA* gene encodes a novel signal transduction protein required for multicellular development. *J. Bacteriol.* 177: 2014-20.

Primakoff, P., and Artz, S. 1979. Positive control of lac operon expression in vitro by guanosine 5'-diphosphate, 3'-diphosphate. *Proc. Natl. Acad. Sci. U.S.A.* 76:1726-1730.

Quillet L, Barray S, Labedan B, Petit F, Guespin-Michel J. 1995. The gene encoding the beta-1,4-endoglucanase (CelA) from *Myxococcus xanthus*: evidence for independent acquisition by horizontal transfer of binding and catalytic domains from actinomycetes. *Gene*. 158(1):23-9.

Ramaswamy S, Dworkin M, Downard J. 1997. Identification and characterization of *Myxococcus xanthus* mutants deficient in calcofluor white binding. *J. Bacteriol.* 179(9):2863-71.

Raivio TL, Silhavy TJ. 2001. Periplasmic stress and ECF sigma factors. *Annu. Rev. Microbiol.* 55:591-624.

Rao NN, Liu S, Kornberg A. 1998. Inorganic polyphosphate in *Escherichia coli*: the phosphate regulon and the stringent response. *J. Bacteriol.* 180(8):2186-93.

Reitzer L. 2003. Nitrogen assimilation and global regulation in *Escherichia coli*. *Annu. Rev. Microbiol.* 57:155-76.

Rodriguez-Soto JP, Kaiser D. 1997. Identification and localization of the Tgl protein, which is required for *Myxococcus xanthus* social motility. *J. Bacteriol.* 179(13):4372-81.

Rodriguez-Soto JP, Kaiser D. 1997. The tgl gene: social motility and stimulation in *Myxococcus xanthus*. *J. Bacteriol.* 179(13):4361-71.

Rubin EJ, Akerley BJ, Novik VN, Lampe DJ, Husson RN, Mekalanos JJ. 1999. In vivo transposition of mariner-based elements in enteric bacteria and mycobacteria. *Proc. Natl. Acad. Sci. USA.* 96(4):1645-50.

Ruggeri ZM. 2003. Von Willebrand factor, platelets and endothelial cell interactions. *J. Thromb. Haemost.* 1(7):1335-42.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualisation and annotation. *Bioinformatics.* 16(10):944-5.

Sadowsky, M.J., Tong, Z., De Souza, M., Wackett, L.P. 1998. AtzC is a new member of the amidohydrolase protein superfamily and is homologous to other atrazine-metabolizing enzymes. *J. Bacteriol.* 180:152-8.

Sambrook J, Fritsch EF, Maniatis T. 1989. *Molecular Cloning: A Laboratory Manual*. 2nd edition. Book 1. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Shleeva MO, Bagramyan K, Telkov MV, Mukamolova GV, Young M, Kell DB, Kaprelyants AS. 2002. Formation and resuscitation of "non-culturable" cells of *Rhodococcus rhodochrous* and *Mycobacterium tuberculosis* in prolonged stationary phase. *Microbiology.* 148(Pt 5):1581-91.

- Shimkets L.J.** 1986. Correlation of energy-dependent cell cohesion with social motility in *Myxococcus xanthus*. J. Bacteriol. 166(3):837-841.
- Shimkets, L.J.** 1990. Social and developmental biology of the Myxobacteria. Microbiol. Rev. 54: 473-501.
- Shimkets, L.J. & Asher, S.J.** 1988. Use of recombination techniques to examine the structure of the *csg* locus of *Myxococcus xanthus*. Mol. Gen. Genet. 211:63-71.
- Shimkets, L.J., Gill, R.E., Kaiser, D.** 1983. Developmental cell interactions in *Myxococcus xanthus* and the *spoC* locus. Proc. Natl. Acad. Sci. USA 80:1406-10.
- Shimkets, L.J. & Kaiser, D.** (1982) Induction of coordinated movement of *Myxococcus xanthus* cells. J. Bacteriol. 152:462-70.
- Shleeva MO, Mukamolova GV, Telkov MV, Berezinskaia TL, Syroeshkin AV, Biketov SF, Kaprel'iants AS.** 2003. Formation of nonculturable *Mycobacterium tuberculosis* and their regeneration. Mikrobiologiya. 72(1):76-83.
- Shleeva M, Mukamolova GV, Young M, Williams HD, Kaprelyants AS.** 2004. Formation of 'non-culturable' cells of *Mycobacterium smegmatis* in stationary phase in response to growth under suboptimal conditions and their Rpf-mediated resuscitation. Microbiology. 150(Pt 6):1687-97.
- Simunovic V, Gherardini FC, Shimkets L.J.** 2003. Membrane localization of motility, signaling, and polyketide synthetase proteins in *Myxococcus xanthus*. J. Bacteriol. 185(17):5066-75.
- Singer, M, Kaiser, D.** 1995. Ectopic production of guanosine penta- and tetraphosphate can initiate early developmental gene expression in *Myxococcus xanthus*. Genes Dev. 9(13):1633-44.
- Skerker JM, Berg HC.** 2001. Direct observation of extension and retraction of type IV pili. Proc. Natl. Acad. Sci. USA. 98(12):6901-4.

- Sogaard-Andersen L, Kaiser D.** 1996. C factor, a cell-surface-associated intercellular signaling protein, stimulates the cytoplasmic Frz signal transduction system in *Myxococcus xanthus*. Proc. Natl. Acad. Sci. USA. 93(7):2675-9.
- Steer, MW.** 1977. Differentiation of the tapetum in Avena. I. The cell surface. J. Cell Sci. 25:125-38.
- Stephens, J., Artz, S. and Ames, B.** 1975. Guanosine 5'-diphosphate 3'-diphosphate (ppGpp): positive effector for histidine operon transcription and general signal for amino acid deficiency. Proc. Natl. Acad. Sci. USA. 72: 4389-93.
- Sun H, Shi W.** 2001. Genetic studies of mrp, a locus essential for cellular aggregation and sporulation of *Myxococcus xanthus*. J. Bacteriol. 183(16):4786-95.
- Sweet IR, Cook DL, DeJulio E, Wallen AR, Khalil G, Callis J, Reems J.** 2004. Regulation of ATP/ADP in pancreatic islets. Diabetes. 53(2):401-9.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG.** 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25(24):4876-82.
- Thompson JD, Higgins DG, Gibson TJ.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22(22):4673-80.
- Toal DR, Clifton SW, Roe BA, Downard J.** 1995. The esg locus of *Myxococcus xanthus* encodes the E1 alpha and E1 beta subunits of a branched-chain keto acid dehydrogenase. Mol. Microbiol. 16(2):177-89.
- Tojo N, Inouye S, Komano T.** 1993. The *lonD* gene is homologous to the lon gene encoding an ATP-dependent protease and is essential for the development of *Myxococcus xanthus*. J. Bacteriol. 175(14):4545-9.
- Tokunaga C, Yoshino K, Yonezawa K.** 2004. mTOR integrates amino acid- and energy-

sensing pathways. *Biochem. Biophys. Res. Commun.* 313(2):443-6.

Tufariello JM, Jacobs WR Jr, Chan J. 2004. Individual *Mycobacterium tuberculosis* resuscitation-promoting factor homologues are dispensable for growth *in vitro* and *in vivo*. *Infect. Immun.* 72(1):515-26.

Vojnov AA, Zorreguieta A, Dow JM, Daniels MJ, Dankert MA. 1998. Evidence for a role for the gumB and gumC gene products in the formation of xanthan from its pentasaccharide repeating unit by *Xanthomonas campestris*. *Microbiology*. 144 (Pt 6):1487-93.

Wall D, Kaiser D. 1999. Type IV pili and cell motility. *Mol. Microbiol.* 32(1):1-10.

Wall D, Wu SS, Kaiser D. 1998. Contact stimulation of Tgl and type IV pili in *Myxococcus xanthus*. *J. Bacteriol.* 180(3):759-61.

Wall D, Kolenbrander PE, Kaiser D. 1999. The *Myxococcus xanthus pilQ (sglA)* gene encodes a secretin homolog required for type IV pilus biogenesis, social motility, and development. *J. Bacteriol.* 181(1):24-33.

Wang J, Mu Q. 2003. Soap-HT-BLAST: high throughput BLAST based on Web services. *Bioinformatics*. 19(14):1863-4.

Ward MJ, Lew H, Zusman DR. 2000. Disruption of aldA influences the developmental process in *Myxococcus xanthus*. *J. Bacteriol.* 182(2):546-50.

Watkins JC, Evans RH. 1981. Excitatory amino acid transmitters. *Annu. Rev. Pharmacol. Toxicol.* 21:165-204.

Watson BF, Dworkin M. 1968. Comparative intermediary metabolism of vegetative cells and microcysts of *Myxococcus xanthus*. *J. Bacteriol.* 96(5):1465-73.

Weimer RM, Creighton C, Stassinopoulos A, Youderian P, Hartzell PL. 1998. A chaperone in the HSP70 family controls production of extracellular fibrils in *Myxococcus xanthus*. *J. Bacteriol.* 180(20):5357-68.

- Wolgemuth C, Hoiczky E, Kaiser D, Oster G.** 2002. How Myxobacteria glide. *Curr. Biol.* 12(5):369-77.
- Wu SS, Wu J, Cheng YL, Kaiser D.** 1998. The *pilH* gene encodes an ABC transporter homologue required for type IV pilus biogenesis and social gliding motility in *Myxococcus xanthus*. *Mol. Microbiol.* 29(5):1249-61.
- Wu SS, Wu J, Kaiser D.** 1997. The *Myxococcus xanthus pilT* locus is required for social gliding motility although pili are still produced. *Mol. Microbiol.* 23(1):109-21.
- Xu D, Yang C, Kaplan HB.** 1998. *Myxococcus xanthus sasN* encodes a regulator that prevents developmental gene expression during growth. *J. Bacteriol.* 180(23):6215-23.
- Yang C, Kaplan HB.** 1997. *Myxococcus xanthus sasS* encodes a sensor histidine kinase required for early developmental gene expression. *J. Bacteriol.* 179(24):7759-67.
- Yang Z, Geng Y, Shi W.** 1998. A DnaK homolog in *Myxococcus xanthus* is involved in social motility and fruiting body formation. *J. Bacteriol.* 180(2):218-24.
- Yang Z, Guo D, Bowden MG, Sun H, Tong L, Li Z, Brown AE, Kaplan HB, Shi W.** 2000. The *Myxococcus xanthus wbgB* gene encodes a glycosyltransferase homologue required for lipopolysaccharide O-antigen biosynthesis. *Arch. Microbiol.* 174(6):399-405.
- Yoshida T, Ayabe Y, Yasunaga M, Usami Y, Habe H, Nojiri H, Omori T.** Genes involved in the synthesis of the exopolysaccharide methanolan by the obligate methylophilic *Methylobacillus* sp strain 12S. *Microbiology.* 2003 Feb;149(Pt 2):431-44.
- Yuan J, Amend A, Borkowski J, DeMarco R, Bailey W, Liu Y, Xie G, Blevins R.** 1999. MULTICLUSTAL: a systematic method for surveying Clustal W alignment parameters. *Bioinformatics.* 15(10):862-3.
- Zhang H.** 2003. Alignment of BLAST high-scoring segment pairs based on the longest increasing subsequence algorithm. *Bioinformatics.* 19(11):1391-6.

Zhou Y, Huang GM, Wei L. 2002. UniBLAST: a system to filter, cluster, and display BLAST results and assign unique gene annotation. *Bioinformatics*. 18(9):1268-9.

Zhu W, Plikaytis BB, Shinnick TM. 2003. Resuscitation factors from *Mycobacteria*: homologs of *Micrococcus luteus* proteins. *Tuberculosis (Edinb)*. 83(4):261-9.